Models and Monitoring Designs for Spatio-temporal Climate Data Fields

by

Camila Maria Casquilho Resende

B.Sc. Actuarial Science, Federal University of Rio de Janeiro, 2008
B.Sc. Statistics, Federal University of Rio de Janeiro, 2010
M.Sc. Statistics, Federal University of Rio de Janeiro, 2011

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

The Faculty of Graduate and Postdoctoral Studies

(Statistics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

September 2016

Abstract

In this thesis, we describe how appropriately modelling the spatio-temporal mean surface can help resolve complex patterns of nonstationarity and improve spatial prediction. Nonstationary fields are common in environmental science applications, and even more so in regions with complex terrain. Our analyses focus on the Pacific Northwest, a region where rapid changes and localized weather are very common, and where the terrain plays an important role in separating often radically different climate and weather regimes. To this end, we introduce two comparable strategies for spatial prediction. The first is based on a Bayesian spatial prediction method, where an exploratory analysis was performed in order to better understand the localized weather regimes. The other is based on tackling the anomalies of expected climate in the Pacific Northwest region, based on the average values of temperature computed over a 30-year range obtained through a climate analysis system.

Secondly, we focus on one of the recent challenges in spatial statistics applications, the data fusion problem. There has been an increased need for combining information from multiple sources that may be on different spatial scales. Ensemble modelling is referred to as a statistical post-processing technique based on combining multiple computer model outputs in a statistical model with the goal of obtaining probabilistic forecasts. We give an overview of some ensemble modelling strategies, by combining observed temperature measurements with outputs from an ensemble of deterministic climate models. We also provide a comparison between the Bayesian model averaging approach and a dynamic Bayesian ensemble strategy for forecasting.

Finally, we introduce a novel strategy for the design of monitoring network, where the goal is to select a high-quality yet diverse set of locations. The idea of spatial repulsion is brought to this context via the theory of determinantal point processes. Our design strategy is not only able to yield spatially-balanced designs, but it also has the ability to assess similarity between the potential locations should there be extra sources of information related to the underlying process of interest. We explore its relationship to existing design methods, such as the entropy-based and space-filling designs.

Preface

This thesis is an original work of the author, Camila Maria Casquilho Resende, under the supervision of Dr. James V. Zidek and Dr. Nhu D. Le.

Dr. Alexandre Bouchard-Côté provided valuable suggestions at the early stages of the development of Chapter 8. A version of Chapter 4 was submitted to peer review. The manuscript is called "Spatio-temporal modelling of temperature fields in the Pacific Northwest", by Casquilho-Resende, C. M., Le, N. D. and Zidek, J. V. The idea was jointly developed by myself, Dr. Nhu D. Le and Dr. James V. Zidek. I conducted all computational work, derivations, and the majority of the writing. An electronic version can be found online at arxiv:1604.00572.

Table of Contents

Appendices

List of Tables

List of Figures

Acknowledgments

I am deeply thankful to my supervisors Prof. Jim Zidek and Dr. Nhu Le, for the opportunity of working with them and for their continuous support throughout the development of this thesis.

I would also like to thank Prof. Alexandre Bouchard-Côté for the invaluable discussions about my work and about machine learning (ML) in general. Thanks to the opportunity of being part of the ML reading group meetings. This has been instrumental to the broadening of my research interests.

Thanks to STATMOS/PIMS-UBC for the financial support to attend valuable workshops that greatly contributed to the development of this work.

I am also thankful to my dear friends in UFRJ and UBC who believed in and gave me continuous incentive to carry on this journey, despite my self-criticism. Special thanks goes to my mum, for her unconditional love and support.

Dedication

À minha mãe

Chapter 1

Introduction

Spatial statistics focuses on the modelling of processes where geographical information is of interest or relevant to understand an underlying physical phenomenon. The areas of application of these methodologies are vast, such as in environmental sciences, forestry, and agriculture. It is often essential that we understand these phenomena to better understand their effects in nature. Notably in environmental science, there has been a growing need for understanding the changes in the Earth's climate as well as increasing concerns due to their potential impact on human health. Our work is mainly motivated by these concerns. We focus on diverse objectives related to environmental sciences, which we describe below.

This thesis starts by providing some background on spatial statistics in Chapter 2. In Chapter 3, we provide an overview of approximate methods for performing Bayesian inference.

In Chapter 4, we analyze temperature fields in the Pacific Northwestern region. The importance of modelling temperature fields goes beyond the need to understand a region's climate and serves too as a starting point for understanding their socioeconomic, and health consequences. Particularly due to the topography of this region, temperature modelling has been recognized to be challenging (Mass, 2008; Salathé et al., 2008; Kleiber et al., 2013), and demands flexible spatio-temporal models that are able to handle nonstationarity and changes in trend.

Our main message is on how appropriately modelling the spatio-temporal mean can help resolve complex patterns for nonstationarity and improve spatial prediction. This is often achieved with an exploratory analysis to better understand the localized changes in trend, instead of simply focusing on the modelling of the spatial covariance structure. We argue that carefully mod-

elling the spatio-temportal mean is needed to better represent interesting smaller-scale trends, especially for regions with a complex terrain like the Pacific Northwest, which may not be captured by global climate models. Another contribution is the ability to accommodate features in the mean that vary over space by extending the spatio-temporal model proposed in Le and Zidek (1992). This methodology is flexible and able to accommodate nonstationarity.

We then introduce two comparable strategies for performing spatial prediction. The first is based on the extended Bayesian spatial prediction method after an exploratory analysis to better understand the local changes in trend, and the realization of the need to account for interacted spatio-temporal features in the mean. The second is based on tackling the anomalies of expected climate in the Pacific Northwest, based on the average values of temperature computed over a 30-year range (1981-2010), provided by PRISM Climate Group (Daly et al., 1994, 1997, 2000). For the latter strategy, we observed a higher mean squared prediction error for out-of-the-sample monitoring stations.

Subsequently, in Chapter 5, we explore the data fusion problem, where our goal is to combine information from multiple sources that might have been measured at different spatial scales. In weather studies, data measurements are often supplemented by information brought by computer model outputs. These computer models simulate physical phenomena in order better understand complex physical systems. We provide a description of the Bayesian Ensemble Melding model (BEM) methodology introduced by Liu (2007) following Fuentes and Raftery (2005). The main idea lies in linking processes on mismatched scales through an underlying "true" process. One of the main disadvantages of these methodologies is the computational burden faced while performing inference, as noted in many applications (Swall and Davis, 2006; Smith and Cowles, 2007; Foley and Fuentes, 2008). Here, we introduce a scalable inference methodology alternative for the BEM using integrated nested Laplace approximations (INLA) (Rue et al., 2009). Following Lindgren et al. (2011), we take advantage of a Markov representation of the Matérn covariance family in a continuous space. We illustrate

the methodology for combining an ensemble of computer model outputs with data measurements of temperature across the Pacific Northwest.

Then in Chapter 6, we introduce a dynamic strategy with the objective of performing forecasting that builds on the BEM model's ability to accommodate time. The methodology uses an INLA framework and is computationally efficient. We provide a comparison of the DBEM forecasting strengths with a Bayesian Model Averaging (BMA) (Raftery et al., 2005) alternative. The DBEM methodology is based on a mixture of posterior distributions in a training set over time. Our empirical studies indicate that the DBEM is able take advantage of this smoothing by borrowing strength of nearby sites, yielding less uncertainty in forecasting intervals, but it underperforms the BMA under a lot of uncertainty *a posteriori*.

Afterwards, in Chapter 8 we stress how monitoring networks play an important role in surveillance of environmental processes. We introduce a flexible monitoring network design strategy based on k-determinantal point processes (DPP) (Kulesza and Taskar, 2012). An overview of DPPs is provided in Chapter 7. The k-DPP design is able to yield a spatially balanced design by imposing repulsion on the distances between existing locations and hence avoiding spatial clumping, but also has the ability to assess similarity between the potential locations should there be extra sources of information known to influence the underlying process of interest. We describe how the methodology is able to handle both designing and redesigning of a monitoring network.

Moreover, our empirical studies illustrate how the k-DPP sampling design strategy can be used as a spatially balanced sampling design alternative to the space-filling design (Royle and Nychka, 1998). The main advantage is due to to the fact that it is essentially a randomized design strategy, which can be useful to help mitigate selection bias risks. Due to its flexibility, a sampling k-DPP design strategy is particularly suited to the design of mobile networks. Another notable characteristic of the k-DPP design objective is that it is constructed in such way that is strongly similar to the entropy-based design (Caselton et al., 1992), and can be viewed as a randomized version of this design. We introduce a sampling strategy based on k-DPP

that can be particularly useful to approximate the optimal solution for the entropy-based design when the number of combinations is prohibitive, due to the NP-hardness of this design criterion (Ko et al., 1995).

Finally, Chapter 9 reflects on what we have learned in the work reported in this thesis and in turn the future research to which our work leads.

Chapter 2

Spatial Statistics

Spatial statistics comprises a wide range of statistical methods intended for the analysis of georeferenced data. Rapid advances in technology contribute to the ease of collecting spatially referenced data. In science, spatial data arise in many applications, including environmental sciences, epidemiology, agriculture, and image processing, to name just a few. Notably in environmental science, there is a need for understanding the changes in the Earth's climate as well as their impact in human health. Therefore, spatial statistics studies are not only required from a scientific perspective, but also for regulatory purposes. The spatial statistics literature is often divided into three main branches: geostatistics, lattice data and point patterns (Cressie, 1993; Cressie and Wikle, 2011; Banerjee et al., 2014).

In geostatistical studies, the idea is that there exists an underlying spatial process that governs a particular physical phenomenon, but data are only observed at a finite set of locations. The locations, however, are considered fixed. This theory is often used to understand weather phenomena, such as temperature fields, which will be the focus of Chapter 4. In such applications, a monitoring network refers to the weather stations at which the data are recorded. An overview is provided in Section 2.1.

Furthermore, the analysis of lattice data refers to data that are obtained on subregions that make up a larger space. An example would be pixel values from remote sensing. Despite the terminology, data need not be observed in regularly spaced locations. For instance, in epidemiological studies, there is an interest in mapping occurrences of diseases, but the information is often gathered in a provincial or state scale.

Finally, point patterns are associated with studies in which the main interest lies in the location of event occurrences, and assessing whether there

may be a systematic pattern. Unlike geostatistical studies, there is randomness associated to the locations. A brief overview is provided in Section 2.2.

2.1 Overview of Geostatistics

Geostatistics is a branch of spatial statistics in which inference for a spatially continuous phenomenon is based on spatially discrete sampled data, for instance, measurements of temperature obtained at several different weather stations.

In geostatistics, the interest is in a latent continuous process $\{Y^*(\mathbf{s}) : \mathbf{s} \in \mathbb{R}^d\}$, where d is the dimension of the space. Geostatistical analysis, however, is based on a real-valued process $\{Y(\mathbf{s}) : \mathbf{s} \in \mathcal{G}\}$, which is a partial realization of the process on the whole space, where $\mathcal{G} \subset \mathbb{R}^d$. The data come from measurements at each location $\mathbf{s}_i$, denoted as $Y(\mathbf{s}_i)$, $i = 1, \ldots, n$.

A stochastic process is assumed to be a spatial Gaussian process if, for a finite collection of locations $\mathbf{s}_i$, $i = 1, \ldots, n$, the joint distribution of $\mathbf{Y} = (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))$ is multivariate Gaussian. Gaussian processes have been central in spatial statistics, particularly in geostatistical studies. Gelfand and Schliep (2016) provides a description of how Gaussian processes have become "the most valuable tool in geostatistical modelling". One of the main advantages is that it suffices to describe a mean and covariance structure and, additionally, the marginal and conditional distributions are known.

Having said this, there are other works aimed at addressing scenarios in which the normality assumption is not realistic, such as transformations of the data as in De Oliveira et al. (1997), the flexibility of generalized linear models in a geostatistical framework proposed by Diggle et al. (1998), or even alternative stochastic representations introduced by a scale mixing of a Gaussian process as in Palacios and Steel (2006).

As pointed out in Diggle and Ribeiro Jr (2007), the basic geostatistical

model assumes that for each location $\mathbf{s}$,

$$Y(\mathbf{s}) = \mathbf{x}(\mathbf{s})^\top \boldsymbol{\beta} + \eta(\mathbf{s}) + \epsilon(\mathbf{s}),\ \epsilon(\mathbf{s}) \sim N(0, \tau^2), \tag{2.1}$$

where the mean surface $\mu(\mathbf{s}) = \mathbf{x}(\mathbf{s})^\top \boldsymbol{\beta}$ is a linear function of some spatially-referenced explanatory variables stored in the vector $\mathbf{x}(\mathbf{s})^\top$, $\eta(\mathbf{s})$ is a second-order stationary process with zero mean and variance σ^2, as well as an isotropic correlation function, and $\epsilon(\mathbf{s})$ is an uncorrelated Gaussian process with zero mean and variance τ^2. The variance τ^2 is referred to as the nugget effect and interpreted as measurement error and small-scale variation. The quantity $\tau^2 + \sigma^2$ is known as the sill whereas the variance σ^2 is known as the partial sill. In the above and throughout this thesis, $\top$ denotes transpose.

A random process $\eta(\mathbf{s})$ is second-order stationary if

$$\begin{aligned} \mathrm{E}[\eta(\mathbf{s})] &= \mu(\mathbf{s}) = \mu, \quad \text{and} && (2.2) \\ \mathrm{Cov}(\eta(\mathbf{s}+\mathbf{h}), \eta(\mathbf{s})) &= C(\mathbf{s}_i - \mathbf{s}_j), && (2.3) \end{aligned}$$

that is, $\mu(\mathbf{s})$ is constant for all $\mathbf{s} \in D$ and the covariance between any two points $\mathbf{s}$ and $\mathbf{s}+\mathbf{h}$ in D depends only in the separation vector $\mathbf{h}$. $C(\cdot)$ is known as covariance function.

Furthermore, a random process $\eta(\mathbf{s})$ is intrinsic stationary if

$$\begin{aligned} \mathrm{E}[\eta(\mathbf{s}+\mathbf{h}) - \eta(\mathbf{s})] &= 0, \quad \text{and} && (2.4) \\ \mathrm{Var}[\eta(\mathbf{s}+\mathbf{h}) - \eta(\mathbf{s})] &= 2\gamma(\mathbf{h}), && (2.5) \end{aligned}$$

for all $\mathbf{s}$ and $\mathbf{s}+\mathbf{h} \in D$. The quantity $2\gamma(\cdot)$ is known as a variogram whereas $\gamma(\cdot)$ is known as a semivariogram, which are useful to describe spatial dependency. Note that $\mathrm{Cov}(\eta(\mathbf{s}+\mathbf{h}), \eta(\mathbf{s}))$ can be written as

$$\mathrm{Cov}(\eta(\mathbf{s}+\mathbf{h}), \eta(\mathbf{s})) = \sigma^2 \rho(||\mathbf{h}||), \tag{2.6}$$

where $\rho(||\mathbf{h}||) = \mathrm{Corr}\{\eta(\mathbf{s}), \eta(\mathbf{s}+\mathbf{h})\}$. Hence, another way of defining the

semivariogram is through the correlation function ρ, where

$$\gamma(\mathbf{h}) = \sigma^2(1 - \rho(\mathbf{h})), \tag{2.7}$$

where $\rho(\mathbf{h}) = \text{Corr}\{\eta(\mathbf{s}), \eta(\mathbf{s}+\mathbf{h})\}$.

When the covariance function depends only on the distance between the sites, that is,

$$\text{Cov}(\eta(\mathbf{s}+\mathbf{h}), \eta(\mathbf{s})) = C(||\mathbf{h}||), \tag{2.8}$$

the process is known as isotropic. An isotropic and intrinsic stationary process is known as homogenous process.

In the classical spatial textbooks Cressie (1993); Cressie and Wikle (2011); Diggle and Ribeiro Jr (2007), there are several examples of isotropic parametric covariance functions, such as the Matérn family. In the Matérn family, the covariance function is

$$C(u) = \sigma^2 \left\{2^{\kappa-1}\Gamma(\kappa)\right\}^{-1} (u/\phi)^{\kappa} K_{\kappa}(u/\phi), \tag{2.9}$$

where u is the Euclidean distance between two locations, Γ denotes the gamma function, K_{κ} is the modified Bessel function of order $\kappa > 0$. The parameters σ^2 and $\phi > 0$ are the partial sill and scale, respectively.

The smoothness of the process is governed by the parameter κ. When $\kappa = 0.5$, the Matérn covariance function reduces to the exponential, defined as

$$C(u) = \sigma^2 \exp(u/\phi). \tag{2.10}$$

It should be noted that the isotropy assumption is usually unrealistic in environmental applications and there are numerous works in the literature aimed at handling nonstationarity. An overview of such methods is provided in Section 2.1.1.

One important objective in geostatistics is to perform spatial prediction at unobserved locations. Kriging methods provide the best linear unbiased

estimate of the field at unobserved locations. A detailed description of Kriging methods can be found in the classical spatial textbooks Cressie (1993); Diggle and Ribeiro Jr (2007); Cressie and Wikle (2011). More recently, there has been an increased interest in combining multi-source spatially referenced data. In studies about the weather, for instance, besides the data obtained from monitoring stations, outputs from deterministic climate models could provide additional information regarding large-scale variations about the underlying phenomenon and could ultimately improve spatial prediction. Since the climate model outputs and monitoring data are often on mismatched scales, there has been a growing interest in studying techniques for handling this change of support problem. This will be the central focus of Chapter 5.

Another important objective in geostatistical studies is deciding where to position a monitoring station. The design problem will be the central focus of Chapter 8.

2.1.1 Handling Nonstationarity

In environmental applications, the isotropy assumption is often unrealistic and it is crucial to handle nonstationarity in the spatial modelling. Recently, many techniques have been developed. A simple approach is to consider locally stationary models, based on the idea that the effects of nonstationarity in smaller spatial domains may be negligible. Haas (1990) suggests a moving-window technique, based on a circular subregion to where the inference is restricted. The idea was later extended to a spatio-temporal case (Haas, 1995).

Higdon et al. (1999) propose a model based on a moving average specification of a Gaussian process whereas Fuentes and Smith (2001) and Fuentes (2001, 2002) consider a class of nonstationary processes based on mixture of local stationary processes. Unlike the moving-window approaches, the model is defined on the whole region of interest, though locally it still behaves like a stationary process.

Another idea is to assume that after some deformation of the space, the

process may then be assumed stationary. Of particular note is the Sampson-Guttorp approach (Sampson and Guttorp, 1992) to spatial deformation, which considers models of the form

$$\text{Cov}(Y(\mathbf{s}_1), Y(\mathbf{s}_2)) = 2\rho_\theta(||f(\mathbf{s}_1) - f(\mathbf{s}_2)||), \tag{2.11}$$

where f is a smooth nonlinear map $\mathcal{G} \to \mathcal{D}$ from the geographical $\mathcal{G}$-space ($\mathcal{G} \subset \mathbb{R}^d$) to the deformed $\mathcal{D}$-space ($\mathcal{G} \subset \mathbb{R}^d$).

The locations of the sites in the $\mathcal{D}$-space are obtained via a multidimensional scaling (MDS) algorithm. A mapping of the sites from the $\mathcal{G}$-space into the $\mathcal{D}$-space is obtained through the minimization problem of the following criterion over all monotonic functions δ:

$$\min_{\delta} \frac{\sum_{i<j}[\delta(d_{ij}) - h_{ij}]^2}{\sum_{i<j} h_{ij}^2}, \tag{2.12}$$

where d_{ij} and h_{ij} denote the observed dispersion and the distance between between sites i and j in the $\mathcal{D}$-space, respectively.

Once the locations of the sites are obtained in the $\mathcal{D}$-space, Sampson and Guttorp (1992) use thin-plate splines to obtain a smooth mapping of the sites from the $\mathcal{G}$-space into the $\mathcal{D}$-space and the δ function is replaced by a smooth function g such that $d_{ij} \approx g(h_{ij})$. It is then possible to obtain estimates of realizations of the spatial process at ungauged locations by first smoothly mapping them onto the $\mathcal{D}$-space and subsequently using standard stationary modeling tools.

Damian et al. (2001) extended the Sampson and Guttorp (1992) approach in a Bayesian framework, where the locations in the deformed spaced and the unknown parameters are estimated jointly, thus obtaining a smooth extrapolation of the deformed space to the whole region of interest. Schmidt and O'Hagan (2003) then proposed a similar spatial deformation method based on a Bayesian model, though the mapping of sites is handled in a single framework, and unlike Sampson and Guttorp (1992), their predictive inferences take into account the uncertainty in the mapping. Schmidt and O'Hagan (2003) argue that the Sampson and Guttorp (1992) method

does not account for uncertainty about the mapping since their prediction method is based on some fixed locations in the distorted space.

Other approaches consider the idea of decomposing the covariance function for nonstationary processes, such as the work of Nychka and Saltzman (1998), which considered empirical orthogonal functions, and Nychka et al. (2002), based on a wavelet basis decomposition. More recently, Bornn et al. (2012) proposed a dimension expansion approach, based on the idea that the underlying field can be more straightforwardly described in a higher dimension.

2.2 Overview of Spatial Point Processes

A point process $\mathbf{X}$ is a stochastic mechanism whose realizations consist of countable sets of points, often referred to as events or point patterns. When this process generates a countable set of events in a limited region $D \subset \mathbb{R}^k$, it is called a spatial point process. In practice, the locations where those events occur are of special importance and are modelled as random variables. In particular, the focus is often on understanding and assessing patterns in the locations of these events.

The analysis of spatial point processes can be seen in various scientific applications. For instance, often in forestry applications there may be an interest in determining whether there is a pattern in the locations of a certain specie of trees in a forest, or in monitoring forest wildfires. Another common area where the analysis of spatial point process is found is in epidemiology applications, where the goal could be in monitoring whether there exists a cluster in the locations of occurrence of a certain disease.

Spatial point processes are usually classified based on the pattern of the points. A completely random process is when there is no obvious pattern or structure in the points. These processes are often modelled using a homogenous Poisson process, and are sometimes referred to as complete spatial randomness. The notion of homogeneity is due to the assumption that the number of points falling in a region B is proportional to its area (or volume) $|B|$ on average. More specifically, a homogeneous Poisson process $\mathbf{X}$

in $D \subset \mathbb{R}^k$ with intensity $\lambda > 0$ satisfies the following properties:

- The random variable $Y(B)$ representing the number of events $B \subset D$ follows a Poisson distribution with mean $\mu(B) = \lambda|B|$.
- For non-overlapping regions $B_1, \ldots, B_m$, the number of events in each region are mutually independent random variables.

For more general inhomogenous Poisson process, the random variable $Y(B)$ representing the number of events $B \subset D$ follows a Poisson distribution with mean $\mu(B) = \int_B \lambda(x)dx$.

When a pattern does exist, it may due to a clustering of events, in which case it would be reasonable to assume that the occurrence of an event in a region is associated with occurrence of events nearby. Those processes are often referred to as aggregative point process.

On the other hand, when events are rather evenly spaced, it is reasonable to assume that the occurrence of an event in a region is actually preventing the occurrence of events nearby, that is, repelling events. Those processes are often called repulsive or regular point processes. In Chapter 7, we introduce one process of such type called called determinantal point processes.

In order to describe a spatial point process, we need to define its first and second order properties. The intensity function is defined as

$$\lambda(x) = \lim_{|dx| \to 0} \left\{ \frac{\mathbb{E}[N(dx)]}{|dx|} \right\}, \tag{2.13}$$

where dx is an infinitesimal region that contains the point x, $N(dx)$ denotes the number of events in this infinitesimal region, and $|dx|$ denotes the area or volume of this infinitesimal region.

An estimate of the intensity function can be obtained by assuming a parametric model on the intensity function or by using kernel density estimators (Diggle, 1985; Berman and Diggle, 1989; Bivand et al., 2013). In the

plane,

$$\hat{\lambda}(x) = \frac{1}{h^2} \sum_{j=1}^{N} \frac{M\left(\frac{||x-x_j||}{h}\right)}{q(||x||)}, \tag{2.14}$$

where $x_j \in \{x_1, \ldots, x_N\}$ is an observed point, M is a bivariate symmetric kernel function, and $h > 0$ is the bandwidth controlling the amount of smoothing in the estimation. In practice, however, we only observe points in a window $W \in D$ where the point pattern was observed. The number of points in a circle centred on an point inside the window W is not observable if the circle extends beyond W, which creates an edge effect. Hence, $q(||x||)$ is a border correction to compensate for the missingness due to these edge effects.

The second-order intensity function is defined as

$$\lambda(x_i, x_j) = \lim_{|dx_i|,|dx_j|\to 0} \left\{ \frac{E[N(dx_i)N(dx_j)]}{|dx_i||dx_j|} \right\}. \tag{2.15}$$

where x_i and x_j denote two events in D. The second-order intensity function is related to the chances of any pair of events occurring in the vicinities of x_i and x_j.

When the spatial point process is stationary, its intensity function is constant, i.e. the mean number of events per unit area is $\lambda(x) = \lambda$, and the second-order intensity function is reduced to

$$\lambda_2(x, y) \equiv \lambda_2(x_i - x_j). \tag{2.16}$$

For a stationary, isotropic process the second-order intensity function is reduced to

$$\lambda_2(x, y) \equiv \lambda_2(||x_i - x_j||). \tag{2.17}$$

Another second-order property of a stationary process is given by the

following function

$$K(t) = \lambda^{-1}\mathbb{E}(N_0(t)), \tag{2.18}$$

where $N_0(t)$ is the number of events within distance t of an arbitrary event. Ripley's K is often referred to as slightly modified estimator of the one in Ripley (1988), given by

$$\hat{K}(t) = \frac{1}{n(n-1)}|W|\sum_{i=1}^{n}\sum_{j\neq i} e_{ij} I(d_{ij} \leq t), \tag{2.19}$$

where $|W|$ denotes the area of the observation window, e_{ij} is an edge correction weight. The Ripley's K is a very common exploratory methodology to empirically evaluate inter-point dependencies. It is often compared with the theoretical K based on a Poisson process given by $K(t) = \pi t^2$, and serves as a benchmark for no correlation (Baddeley et al., 2015).

A thorough statistical description o these processes can be found in Ripley (1988); Møller and Waagepetersen (2004); Diggle (2013).

2.3 Lambert Conformal Conic Projection

It is extremely important to take into consideration the curvature of the Earth, especially for large regions such as the ones studied in this thesis. In order to do so, instead of considering the geographical coordinates of the observations, such as latitude and longitude, we obtain their corresponding locations based on a particular cartographic projection called the Lambert Conformal Conic Projection.

As described in Snyder (1987), a cartographic projection is a systematic transformation of the latitudes and longitudes of the observations on the surface of the Earth on a plane. In particular, the Lambert Conformal Conic Projection places a cone over the sphere of the Earth and projects the surface conformally (i.e. preserving angles locally) onto the cone. The scale is true along either one or two standard parallels of latitude.

Geographical coordinates, with λ and ϕ denoting longitude and latitude

respectively, can be transformed into Lambert conformal conic projection coordinates by

$$x = \rho \sin\theta \tag{2.20a}$$
$$y = \rho_0 - \rho\cos\theta, \tag{2.20b}$$

where

$$\rho = \frac{R \times F}{\tan^n(\pi/4 + \phi/2)} \tag{2.21}$$
$$\theta = n(\lambda - \lambda_0) \tag{2.22}$$
$$\rho_0 = \frac{R \times F}{\tan^n(\pi/4 + \phi_0/2)} \tag{2.23}$$
$$F = \cos\phi_1 \tan^n(\pi/4 + \phi_1/2)/n \tag{2.24}$$
$$n = \frac{\ln(\cos\phi_1/\cos\phi_2)}{\ln[\tan(\phi/4 + \phi/2)/\tan(\phi/4 + \phi_1/2)]}, \tag{2.25}$$

ϕ_0 and λ_0 denoting the reference latitude and longitude, R the radius of the Earth, ϕ_1 and ϕ_2 the standard parallels. If only one standard parallel is used, i.e. $\phi_1 = \phi_2$, then $n = \sin(\phi_1)$.

Chapter 3

Approximate Bayesian Inference

Let $\mathbf{Y}$ be a random vector with density function or probability mass function given by $p(\mathbf{Y}|\mathbf{\Psi})$, where $\mathbf{\Psi}$ is a parameter vector characterizing the distribution of $\mathbf{Y}$. In a Bayesian framework, before observing data, a probability distribution is assumed for $\mathbf{\Psi}$, namely a prior distribution. This prior distribution refers to the initial uncertainty about $\mathbf{\Psi}$. After observing realizations of $\mathbf{Y}$, namely the data $\mathbf{y}$, and via the Bayes' theorem, one can obtain the posterior distribution of $\mathbf{\Psi}$ as follows

$$p(\mathbf{\Psi}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{\Psi})p(\mathbf{\Psi})}{\int p(\mathbf{y}|\mathbf{\Psi})p(\mathbf{\Psi})d\mathbf{\Psi}}. \tag{3.1}$$

Statistical inference in a Bayesian framework is based on the posterior distribution of $\mathbf{\Psi}$, which contains all probabilistic information about $\mathbf{\Psi}$. In particular, suppose that $\mathbf{\Psi} = (\Psi_1, \ldots, \Psi_k)^\top$. Marginal posterior distributions $p(\mathbf{\Psi_I})$, where $\mathbf{I} \subseteq \{1, \ldots, k\}$, are obtained by

$$p(\mathbf{\Psi_I}|\mathbf{y}) = \int p(\mathbf{\Psi}|\mathbf{y})d\mathbf{\Psi}_{\bar{\mathbf{I}}}, \tag{3.2}$$

where $\bar{\mathbf{I}}$ denote the complement of $\mathbf{I}$.

A common challenge in Bayesian inference is that often it is not possible to analytically solve the integral in equations (3.1) or (3.2). Numerical approximations have been developed mostly in the 1980s, such as Naylor and Smith (1982), Tierney and Kadane (1986), Smith et al. (1987) and Tierney et al. (1989). With the recent advances in computational methods, and perhaps motivated by the work of Gelfand and Smith (1990), Markov

chain Monte Carlo (MCMC) methods have been increasingly popular since the 1990s.

However, for the class of latent Gaussian models, MCMC strategies provide a significant computational burden and the need to deal with the common issue of correlated parameters. In this thesis, we focus on the more recent and well-established work of Rue et al. (2009), that provides a deterministic attractive alternative to the computationally intensive MCMC methods. We describe this methodology in Section 3.2. In the following Section 3.1, we provide an overview of the Laplace's method.

3.1 Laplace's Method

In order to approximate unimodal posteriors, Tierney and Kadane (1986) proposed the use of Laplace method to obtain moments of smooth positive functions. For instance, let $\mathcal{G}$ be a smooth positive function on a parameter space. The posterior mean of $\mathcal{G}(\boldsymbol{\Psi})$ can be written as

$$E[\mathcal{G}(\boldsymbol{\Psi})] = \int \mathcal{G}(\boldsymbol{\Psi}) p(\boldsymbol{\Psi}|\mathbf{y}) d\boldsymbol{\Psi} = \frac{\int \mathcal{G}(\boldsymbol{\Psi}) l(\boldsymbol{\Psi};\mathbf{y}) p(\boldsymbol{\Psi}) d\boldsymbol{\Psi}}{\int l(\boldsymbol{\Psi};\mathbf{y}) p(\boldsymbol{\Psi}) d\boldsymbol{\Psi}}, \tag{3.3}$$

where $p(\boldsymbol{\Psi}|\mathbf{y})$ denotes the posterior distribution of $\boldsymbol{\Psi}$, $l(\boldsymbol{\Psi};\mathbf{y})$ the likelihood function, and $p(\boldsymbol{\Psi})$ the prior distribution of $\boldsymbol{\Psi}$.

Firstly, let $\mathcal{L}(\boldsymbol{\Psi}) = \log(\mathcal{G}(\boldsymbol{\Psi}) l(\boldsymbol{\Psi};\mathbf{y}) p(\boldsymbol{\Psi}))$ denote the logarithm of the integrand of the numerator in equation 3.3. Expanding $\mathcal{L}(\boldsymbol{\Psi})$ around its mode $\boldsymbol{\Psi}^*$ gives

$$\mathcal{L}(\boldsymbol{\Psi}) \approx \mathcal{L}(\boldsymbol{\Psi}^*) - \frac{1}{2}(\boldsymbol{\Psi} - \boldsymbol{\Psi}^*)^\top \mathcal{I}^{-1}(\boldsymbol{\Psi}^*)(\boldsymbol{\Psi} - \boldsymbol{\Psi}^*), \tag{3.4}$$

where $\mathcal{I}^{-1}(\boldsymbol{\Psi}^*) = -[\nabla^2 \mathcal{L}(\boldsymbol{\Psi})]^{-1}$ is minus of the inverse of the Hessian matrix of $\mathcal{L}(\boldsymbol{\Psi})$ evaluated at its mode $\boldsymbol{\Psi}^*$. Hence,

$$\begin{aligned} \int g(\boldsymbol{\Psi}) l(\boldsymbol{\Psi};\mathbf{y}) p(\boldsymbol{\Psi}) d\boldsymbol{\Psi} &\approx \int \exp\left\{ \mathcal{L}(\boldsymbol{\Psi}^*) - \frac{1}{2}(\boldsymbol{\Psi} - \boldsymbol{\Psi}^*)^\top \mathcal{I}^{-1}(\boldsymbol{\Psi}^*)(\boldsymbol{\Psi} - \boldsymbol{\Psi}^*) \right\} d\boldsymbol{\Psi} \\ &= (2\pi)^{-\frac{m}{2}} |\mathcal{I}(\boldsymbol{\Psi}^*)|^{\frac{1}{2}} \exp\left\{ \mathcal{L}(\boldsymbol{\Psi}^*) \right\}, \end{aligned} \tag{3.5}$$

where m is the dimension of the parameter vector $\boldsymbol{\Psi}$.

Similarly, let $\widetilde{\mathcal{L}}(\boldsymbol{\Psi}) = \log(l(\boldsymbol{\Psi};\mathbf{y})p(\boldsymbol{\Psi}))$ denote the logarithm of the integrand of the denominator in equation 3.3. Expanding $\widetilde{\mathcal{L}}(\boldsymbol{\Psi})$ around its mode $\widetilde{\boldsymbol{\Psi}}^*$ gives

$$\widetilde{\mathcal{L}}(\boldsymbol{\Psi}) \approx \widetilde{\mathcal{L}}(\widetilde{\boldsymbol{\Psi}}^*) - \frac{1}{2}(\boldsymbol{\Psi} - \widetilde{\boldsymbol{\Psi}}^*)^\top \widetilde{\mathcal{I}}^{-1}(\widetilde{\boldsymbol{\Psi}}^*)(\boldsymbol{\Psi} - \widetilde{\boldsymbol{\Psi}}^*), \tag{3.6}$$

where $\widetilde{\mathcal{I}}^{-1}(\widetilde{\boldsymbol{\Psi}}^*) = -[\nabla^2 \mathcal{L}(\boldsymbol{\Psi})]^{-1}$ is minus of the inverse of the Hessian matrix $\widetilde{\mathcal{L}}(\boldsymbol{\Psi})$ evaluated at its mode $\widetilde{\boldsymbol{\Psi}}^*$. Hence,

$$\begin{aligned} \int l(\boldsymbol{\Psi};\mathbf{y})p(\boldsymbol{\Psi})d\boldsymbol{\Psi} &\approx \int \exp\left\{\widetilde{\mathcal{L}}(\widetilde{\boldsymbol{\Psi}}^*) - \frac{1}{2}(\boldsymbol{\Psi} - \widetilde{\boldsymbol{\Psi}}^*)^\top \widetilde{\mathcal{I}}^{-1}(\widetilde{\boldsymbol{\Psi}}^*)(\boldsymbol{\Psi} - \widetilde{\boldsymbol{\Psi}}^*)\right\} d\boldsymbol{\Psi} \\ &= (2\pi)^{-\frac{m}{2}} |\widetilde{\mathcal{I}}^{-1}(\boldsymbol{\Psi}^*)|^{\frac{1}{2}} \exp\left\{\widetilde{\mathcal{L}}(\widetilde{\boldsymbol{\Psi}}^*)\right\}. \end{aligned} \tag{3.7}$$

Finally, equation 3.3 can be approximated as

$$E[\mathcal{G}(\boldsymbol{\Psi})] \approx \frac{|\mathcal{I}^{-1}(\boldsymbol{\Psi}^*)|^{\frac{1}{2}}}{|\widetilde{\mathcal{I}}^{-1}(\boldsymbol{\Psi}^*)|^{\frac{1}{2}}} \exp\left\{\mathcal{L}(\boldsymbol{\Psi}^*) - \widetilde{\mathcal{L}}(\widetilde{\boldsymbol{\Psi}}^*)\right\}. \tag{3.8}$$

3.2 Integrated Nested Laplace Approximation

The integrated nested Laplace approximation (INLA) method was proposed by Rue et al. (2009) as a way of performing approximate Bayesian inference for latent Gaussian models. The `INLA` package for computation in `R` is available for download at www.r-inla.org (last accessed on June 15, 2016). In this section, we provide an overview of the INLA method.

Consider the following hierarchical structure

$$y_i|\mathbf{x},\boldsymbol{\theta} \sim p(y_i|x_i,\boldsymbol{\theta}) \tag{3.9a}$$

$$\mathbf{x}|\boldsymbol{\theta} \sim p(\mathbf{x}|\boldsymbol{\theta}) \tag{3.9b}$$

$$\boldsymbol{\theta} \sim p(\boldsymbol{\theta}), \tag{3.9c}$$

where $\mathbf{y} = (y_1, \ldots, y_{n_d})'$ denotes the observed data, $\mathbf{x} = (x_1, \ldots, x_n)'$ the elements of the latent field, $\boldsymbol{\theta}$ a m-dimensional hyperparameter vector, where $y_i|\mathbf{x},\boldsymbol{\theta}$ belongs to the exponential family of distributions and $\mathbf{x}|\boldsymbol{\theta}$ is assumed

to follow a Gaussian distribution. Furthermore, it is assumed that data are conditionally independent given the latent field $\mathbf{x}$. Hence,

$$p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i|x_i, \boldsymbol{\theta}). \tag{3.10}$$

The posterior distribution is then given by

$$p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}) \propto p(\boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\theta}) \prod_{i=1}^{n} p(y_i|x_i, \boldsymbol{\theta}). \tag{3.11}$$

The latent field $\mathbf{x}$ is often of high dimension, but it is assumed that they have conditional independence properties. In particular, Rue et al. (2009) consider models that satisfy the Markov property, i.e, $x_i|\mathbf{x}_{-i}$ only depends on a subset $\mathbf{x}_{-i}$, where $\mathbf{x}_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$. In this case, the latent field is a Gaussian random field and its precision matrix is sparse, which is the key ingredient for computational efficiency.

In particular, Rue et al. (2009) focus on approximating the following marginal distributions

$$p(x_i|\mathbf{y}) = \int p(\boldsymbol{\theta}|\mathbf{y})p(x_i|\boldsymbol{\theta}, \mathbf{y})d\boldsymbol{\theta} \tag{3.12}$$

$$p(\theta_j|\mathbf{y}) = \int p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}_{-j}, \tag{3.13}$$

for every $i = 1, \ldots, n$ and $j = 1, \ldots, m$.

When the integrals in (3.12) and (3.13) cannot be found analytically, approximations for $p(x_i|\boldsymbol{\theta}, \mathbf{y})$ and $p(\boldsymbol{\theta}|\mathbf{y})$ are obtained and denoted by $\tilde{p}(x_i|\boldsymbol{\theta}, \mathbf{y})$ and $\tilde{p}(\boldsymbol{\theta}|\mathbf{y})$, respectively. Thus, one can construct the following nested approximations

$$\tilde{p}(x_i|\mathbf{y}) = \int \tilde{p}(\boldsymbol{\theta}|\mathbf{y})\tilde{p}(x_i|\boldsymbol{\theta}, \mathbf{y})d\boldsymbol{\theta} \tag{3.14}$$

$$\tilde{p}(\theta_j|\mathbf{y}) = \int \tilde{p}(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}_{-j}. \tag{3.15}$$

Via numerical integration, an approximation can be obtained for $\tilde{p}(x_i|\mathbf{y})$

as follows

$$\tilde{p}(x_i|\mathbf{y}) \approx \sum_j \tilde{p}(x_i|\boldsymbol{\theta}_j, \mathbf{y})\tilde{p}(\boldsymbol{\theta}_j|\mathbf{y})\Delta_j \tag{3.16}$$

where Δ_j denotes the area weights associated with the evaluation points $\boldsymbol{\theta}_j$.

Below, we discuss strategies for approximating $p(\boldsymbol{\theta}|\mathbf{y})$, $p(x_i|\boldsymbol{\theta}, \mathbf{y})$, and for the marginal likelihood of the data, $p(\mathbf{y})$.

Approximation of $p(\boldsymbol{\theta}|\mathbf{y})$

The approximation of $p(\boldsymbol{\theta}|\mathbf{y})$ is equivalent to the Laplace approximation for the marginal posterior distribution originally proposed by Tierney and Kadane (1986), and is given by

$$\tilde{p}(\boldsymbol{\theta}|\mathbf{y}) \propto \left.\frac{p(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y})}{\tilde{p}_G(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})}\right|_{\mathbf{x}=\mathbf{x}^*(\boldsymbol{\theta})}, \tag{3.17}$$

where $\tilde{p}_G(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ is the Gaussian approximation of the full conditional distribution of $\mathbf{x}$ and $\mathbf{x}^*(\boldsymbol{\theta})$ is the mode of the full conditional $\mathbf{x}$ for a given $\boldsymbol{\theta}$. The proportional sign in (3.17) is due to the fact that the normalizing constant of $p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})$ is unknown.

Note that

$$p(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y}) \propto \exp\left\{-\frac{1}{2}\mathbf{x}'\mathbf{Q}(\boldsymbol{\theta})\mathbf{x} + \sum_i \log(p(y_i|x_i, \boldsymbol{\theta}))\right\}, \tag{3.18}$$

where $\mathbf{Q}(\boldsymbol{\theta})$ denotes the precision matrix of the Gaussian latent field with hyperparameters $\boldsymbol{\theta}$. Obtaining the Taylor expansion of second order about $\mathbf{x}^*$ gives

$$\tilde{p}(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y}) \propto \exp\left\{-\frac{1}{2}\mathbf{x}'(\mathbf{Q}(\boldsymbol{\theta}) + \text{diag}(\mathbf{c}))\mathbf{x} + \mathbf{d}'\mathbf{x}\right\}, \tag{3.19}$$

where $\mathbf{c}$ and $\mathbf{d}$ are the coefficients of the expansion, and $\text{diag}(\cdot)$ represents a diagonal matrix.

In practice, Rue et al. (2009) note that there is no need to represent

$\tilde{p}(\boldsymbol{\theta}|\mathbf{y})$ parametrically, rather, explore it sufficiently well in order to select "good" points for the numerical integration. That is done by locating the mode $\boldsymbol{\theta}^*$ of $p(\boldsymbol{\theta}|\mathbf{y})$ and exploring the log-posterior in order to select points in regions of high probability mass for use in the integration.

However, should the number of hyperparameters be large, say greater than 5, this grid exploration strategy can be very inefficient, with a computational cost that grows exponentially with the number of hyperparameters. For such cases, Rue et al. (2009) propose the use of a central composite design (CCD) strategy to help locate the integration points, which can then be used to estimate the curvature of $p(\boldsymbol{\theta}|\mathbf{y})$ around its mode.

Marginal Likelihood Approximation

An approximation for the marginal likelihood can be obtained from (3.17),

$$\tilde{p}(\mathbf{y}) = \int \left. \frac{p(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y})}{\tilde{p}_G(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})} \right|_{\mathbf{x}=\mathbf{x}^*(\boldsymbol{\theta})} d\boldsymbol{\theta}, \tag{3.20}$$

where $p(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y}) = p(\boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\theta})p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$.

Approximation of $p(x_i|\boldsymbol{\theta}, \mathbf{y})$

There are different ways of approximating $p(x_i|\boldsymbol{\theta}, \mathbf{y})$ (Rue et al., 2009) as described below. Throughout the applications in this thesis, we opted for a Laplace approximation.

- **Gaussian approximation**: The simplest method of approximating $p(x_i|\boldsymbol{\theta}, \mathbf{y})$ is based on a Gaussian approximation

$$\tilde{p}_G(x_i|\boldsymbol{\theta}, \mathbf{y}) \equiv N(\mu_i(\boldsymbol{\theta}), \sigma_i^2(\boldsymbol{\theta})), \tag{3.21}$$

  where $\mu_i(\boldsymbol{\theta})$ and $\sigma_i^2(\boldsymbol{\theta})$ are the marginal mean and variance from $\tilde{p}_G(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$. This approximation can be useful in some cases, but as noted in Rue and Martino (2007), it may lead to errors in the location or lack of skewness.

- **Laplace approximation**: Another way of approximating $p(x_i|\boldsymbol{\theta}, \mathbf{y})$ is based on a Laplace approximation:

$$\tilde{p}_{LA}(x_i|\boldsymbol{\theta}, \mathbf{y}) \propto \left. \frac{p(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y})}{\tilde{p}_{GG}(\mathbf{x}_{-i}|x_i, \boldsymbol{\theta}, \mathbf{y})} \right|_{\mathbf{x}_{-i}=\mathbf{x}^*_{-i}(x_i, \boldsymbol{\theta})}, \tag{3.22}$$

  where $\mathbf{x}^*_{-i}(x_i, \boldsymbol{\theta})$ is the mode of $p(\mathbf{x}_{-i}|x_i, \boldsymbol{\theta}, \mathbf{y})$ and $\tilde{p}_{GG}(\mathbf{x}_{-i}|x_i, \boldsymbol{\theta}, \mathbf{y})$ denotes the Gaussian approximation for $\mathbf{x}_{-i}|x_i, \boldsymbol{\theta}, \mathbf{y}$. A drawback is the need to evaluate $\tilde{p}_{GG}$ for every x_i and $\boldsymbol{\theta}$, which is computationally inefficient. A way of overcoming this is by avoiding the optimization step and instead obtaining an approximate mode as follows

$$\mathbf{x}^*_{-i}(x_i, \boldsymbol{\theta}) \approx E_{\tilde{p}_G}(\mathbf{x}_{-i}|x_i). \tag{3.23}$$

- **Simplified Laplace approximation**: Rue et al. (2009) introduced yet another way of approximating $p(x_i|\boldsymbol{\theta}, \mathbf{y})$ based on expanding (3.22) about $x_i = \mu_i(\boldsymbol{\theta})$ up to third order, which allows for correcting the Gaussian approximation $\tilde{p}_G(x_i|\boldsymbol{\theta}, \mathbf{y})$ in terms of location and skewness in a more computationally efficient way.

Chapter 4

Temperature Fields in the Pacific Northwest

4.1 Motivation

Meteorological variables are crucial to understand a region's climate. In particular, a much discussed topic in recent years is that the Earth's climate has been changing: global average atmospheric and sea surface temperature have increased and extreme temperature events such as heat waves are now more frequent. This changing climate has led to concerns about its impact on human health.

Extreme temperatures may contribute to cardiovascular and respiratory diseases, especially among elderly people, as is outlined in Åström et al. (2013). Li et al. (2012) studied the relationship between temperature and morbidity due to extreme heat and revealed that a number of hospital admissions in Milwaukee, Wisconsin were detected to be significantly related to high temperature. In fact, Robine et al. (2008) estimates an excess death toll of 70,000 people due to high temperatures in Europe in 2003 and a World Health Organization (WHO) assessment concluded that the modest warming that has occurred since the 1970s was already causing over 140,000 excess deaths annually by the year 2004 (World Health Organization, 2009). The spread of infectious diseases is also now being linked to the climate change as per Hoberg and Brooks (2015). All of this highlights that the importance of modelling temperature fields goes well beyond the natural sciences.

Due to the topography of the study region, the modelling of tempera-

ture fields can be particularly challenging. Kleiber et al. (2013) recognized the difficulty faced by statistical models in capturing complex spatial variability. By analyzing data from the state of Colorado, Kleiber et al. (2013) developed a bivariate stochastic temperature model for minimum and maximum temperature via a nonparametric approach. In the Pacific Northwest, Salathé et al. (2008) focused on the development of a regional climate model run at a 15-km grid spacing. The topography of the study region contributes much to the complexity of modelling these fields and demands flexible spatio-temporal models that are able to handle nonstationarity and changes in trend.

4.1.1 Contributions

One of our contributions is the ability to accommodate features in the mean that vary over space by extending the spatio-temporal model proposed in Le and Zidek (1992), and easily performing spatial prediction. This methodology is described in Section 4.4. Another important feature is its flexibility due to the fact that no structure is assumed for the spatial covariance matrix. The method thus is able to accommodate nonstationarity. We illustrate our analysis based on the Sampson and Guttorp (1992) method for estimating nonstationary spatial covariance structures.

Additionally, our work conclusively shows how appropriately modelling the spatio-temporal mean field can resolve complex patterns for nonstationarity and improve spatial prediction. To this end, we also introduce two comparable strategies for spatial prediction. The first is based on the extended Bayesian spatial prediction method after a thorough exploratory analysis to better understand the local changes in trend, and the need to account for spatio-temporal interactions in the mean. The second is based on tackling the anomalies of expected climate in the Pacific Northwest, based on the average values of temperature computed over a 30-year range (1981-2010), provided by PRISM Climate Group. These data were obtained using a climate model called PRISM (Parameter-elevation Relationship on Independent Slopes Model), described in Daly et al. (1994, 1997, 2000). We

provide an overview of the PRISM in Section 4.3.3.

The outline of this chapter is as follows. We begin by providing a description about the Pacific Northwestern region in Section 4.2, followed by a description of the data sets in Section 4.3. Then, we introduce the Bayesian spatial prediction methodology in Section 4.4. Finally, we discuss the results in Section 4.5

4.2 The Pacific Northwest

The Pacific Northwest is the region in the western part of North America adjacent to the Northeast Pacific Ocean. It is a rather diverse region, with four mountain ranges dominating it, including the Cascade Range, the Olympic Mountains, the Coast Mountains and parts of the Rocky Mountains.

This region is known to have a wet and cool climate overall, though in more inland areas, the climate can be fairly dry, with warmer summers and harsher winters. According to Mass (2008), the Northwest weather and climate are dominated mainly by the Pacific Ocean to the west and the region's mountain ranges that block and deflect low-level air. The ocean moderates the air temperatures year-round and serves as a source of moisture, and the mountains modify precipitation patterns and prevent the entrance of wintertime cold-air from the continental interior.

The terrain is another key element to understand the Pacific Northwest weather. East of the Rocky Mountains is where the coldest air is usually located, but the Rockies preclude this cold air from reaching the Northwest and the the cold air that does manage to cross gets warmer when descending to eastern Washington, Oregon, Cascade Range, British Columbia, and Alaska. The temperatures in this region are thus controlled by the atmospheric circulation patterns, the proximity to the Pacific Ocean and by elevation.

In the literature, spatial modelling in this region is recognized to be rather complex and it has been the subject of critical observation by local weather scientists. More recently, Mass (2008) apprises that the weather in

the Northwest is often surprising, both in its intensity and in the remarkable contrasts between nearby locations. Rapid changes and localized weather are very common in this region and the terrain plays an important role in separating often radically different climate and weather regimes. Mote (2004) noticed an apparent tendency for high-elevation stations to exhibit weaker warming trends than lower-elevation stations when examining temperature trends in this region. In order to better understand this region, Salathé et al. (2008) implies that global simulations may indicate large-scale patterns of change though they may not capture the effects of narrow mountain ranges.

Increases in temperature have been observed throughout the Northwest. Across the region from 1895 to 2011, a regionally averaged warming of about 0.7 degrees Celsius (1.3 Fahrenheit) was observed (Kunkel et al., 2013; Mote et al., 2014). This change in climate influences hydrological and biological changes, and ultimately, may lead to economic and social consequences. All of this shows the importance of understanding the temperature fields in this region.

4.3 Data Description

In this section we briefly describe the various temperature data sets considered in this thesis and their sources. Most map images displayed in this thesis were obtained via the `ggmap R` package (R Core Team, 2014; Kahle and Wickham, 2013).

4.3.1 University of Washington (UW) Probcast Group Data

The Probcast data set includes forecasts of surface level temperature data 48 hours ahead, initialized at midnight Coordinated Universal Time (UTC). Data are available for download at the University of Washington (UW) Probcast Group web page (`http://www.stat.washington.edu/MURI/`, last accessed on June 15, 2016.). One of Probcast Group project's main goals was to create methods for the integration of multisource infor-

mation, derived from deterministic model outputs, observations, and expert knowledge.

The time range availability for the Probcast data is from January 12, 2000 to June 30, 2000. This six-month period have become central for this thesis. Stations whose types refer to ship reports, test automated surface observing system reports from the National Weather Service (NWS) or unidentified were considered unreliable by the Probcast Group and thus have not been considered in this analysis. Also, we did not consider fixed buoys type of stations. Figure 5.1 contains a map with the 833 monitoring stations, spread over the Pacific Northwest.

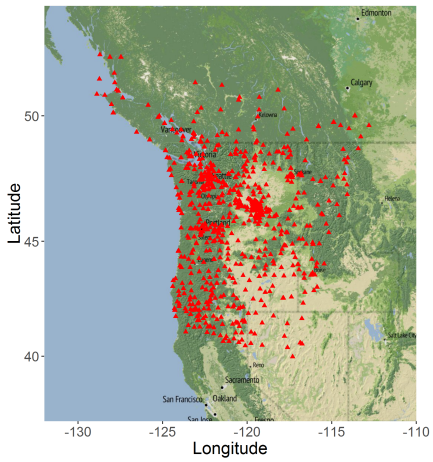


Figure 4.1: Map of the 833 monitoring stations. Map created using `R` library `ggmap` with tiles by Stamen Design, under CC BY 3.0, and data by OpenStreetMap, under CC BY SA.

The data come from the UW mesoscale short-range ensemble system for the Pacific northwestern area, described in detail in Grimit and Mass (2002). It corresponds to a five-member short-range ensemble consisting of different runs of the Pennsylvania State University–National Center for Atmospheric Research fifth generation Mesoscale Model (MM5). The runs differ due to the different initial values considered. In these data, the grid-scaled deterministic model outputs have been interpolated to the locations of the monitoring stations by the Probcast group.

The MURI data set is not an integrated spatio-temporal data set in the sense that the locations in which the measurements are available may vary considerably for different days whilst some stations have very few measurements and model outputs available. The temporal spacing of the observations is highly irregular as it can be seen in Figure 4.2.

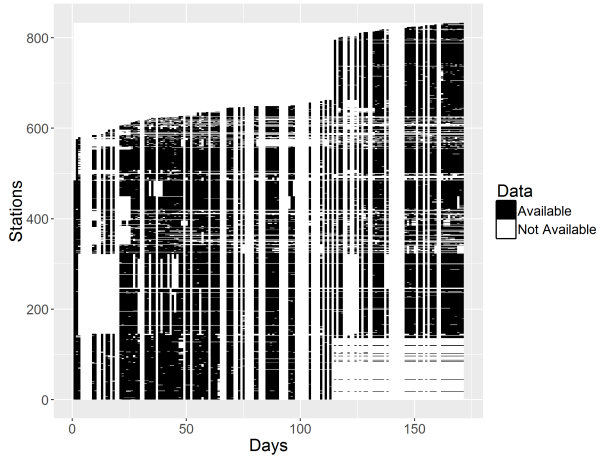


Figure 4.2: Data availability. Notice the unsystematic missingness of the data pattern.

4.3.2 U.S. Global Historical Climatology Network

The U.S. Global Historical Climatology Network - Daily (GHCND) is an integrated database of climate summaries from land surface stations across the globe, developed for several potential applications, including climate analysis and studies that require data at a daily time resolution, as described in Menne et al. (2012) and Lawrimore et al. (2011).

Figure 4.3 illustrates the locations of 97 stations where maximum daily temperature data were downloaded from the GHCND database for the investigation reported in this thesis. Due to the irregular temporal spacing of the observations as discussed in Section 4.3.1, we also collect the GHCND data. Our goal is to emulate the 48-hour forecasts of surface level temperature data from the Probcast group, hence for illustration purposes, the selected spatio-temporal data time frame is from January to June of the

year 2000.

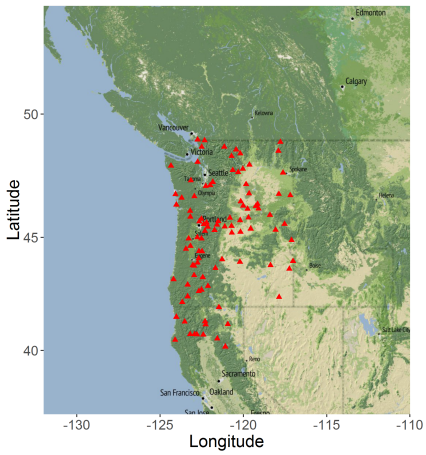


Figure 4.3: Locations of 97 stations in the Pacific Northwestern area considered in this study. Map created using `R` library `ggmap` with tiles by Stamen Design, under CC BY 3.0, and data by OpenStreetMap, under CC BY SA.

Figure 4.4 shows contours of average site temperatures for different months, obtained by bivariate linear interpolation. Notice that cooler temperatures are observed closer to the Pacific Ocean. Another interesting feature is the different patterns of temperature variation across the region. Warmer temperatures are generally found east of the Cascades and since western Washington is more exposed to air coming from from Puget Sound, the Straits of Juan de Fuca and Georgia, and the Pacific Ocean, it generally experiences cooler temperatures.

Our preliminary analysis starts with an exploration of the spatio-temporal trend, followed by an analysis of the unexplained residuals of the spatio-temporal process.

Initially a simple geostatistical model was considered where the spatial trend is described through a second-order polynomial regression model. The temperature measured on day t at location s is denoted as $Y_t(s)$, where t denotes the time in days and $\mathbf{s} = (s_1; s_2)$, the location coordinates, in km, after a suitable projection of the relevant part of the globe onto a flat surface. We considered projected spatial coordinates using the Lambert

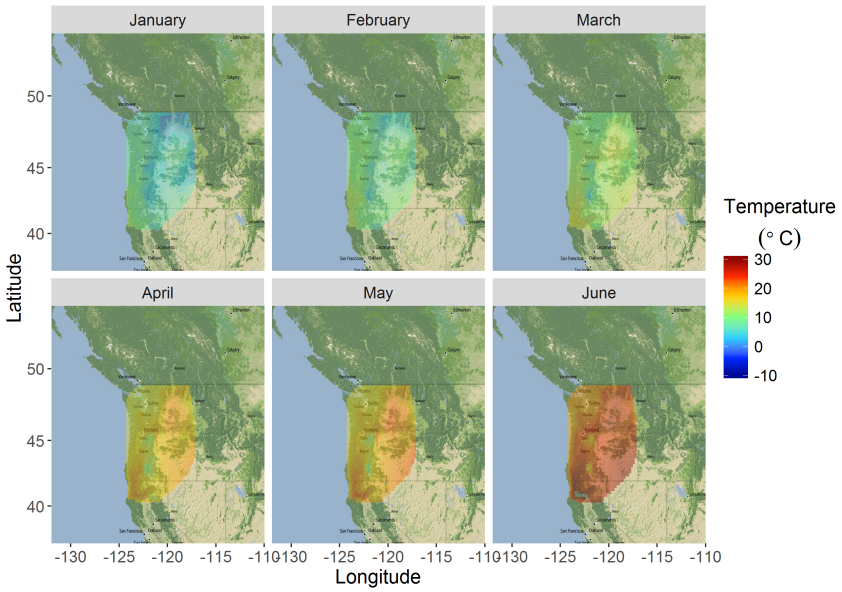


Figure 4.4: Averaged site temperatures for different months. Notice the different patterns of temperature variation across the region. Map created using `R` library `ggmap` with tiles by Stamen Design, under CC BY 3.0, and data by OpenStreetMap, under CC BY SA.

conformal conic projection, but for simplicity, we still refer to these projected coordinates as simply latitude and longitude.

For a fixed time t,

$$\begin{aligned} Y_t(\mathbf{s}) &= \mu_t(\mathbf{s}) + \nu_t(\mathbf{s}),\ \nu_t(\mathbf{s}) \sim N(0, \sigma^2) \\ \mu_t(\mathbf{s}) &= \alpha_t + \beta_{1t}s_1 + \beta_{2t}s_2 + \beta_{3t}s_1s_2 + \beta_{4t}s_1^2 + \beta_{5t}s_2^2. \end{aligned} \tag{4.1}$$

Recall that when the spatial random field is stationary, the semivariogram between two locations $\mathbf{s}_k$ and $\mathbf{s}_l$ at a fixed time point t is defined as

$$\gamma(\mathbf{s}_k, \mathbf{s}_l) = \frac{1}{2}\mathbb{E}[(Y_t(\mathbf{s}_k) - Y_t(\mathbf{s}_l))^2]. \tag{4.2}$$

For illustrative purposes, Figure 4.5 contains the binned empirical semi-

variogram obtained separately for each of the two selected days (January 04 and June 21). The shaded area corresponds to Monte Carlo envelopes obtained by repeatedly recomputing and plotting the semivariance after permutations of the temperature data across the sampling locations. They indicate regions of uncorrelated data.

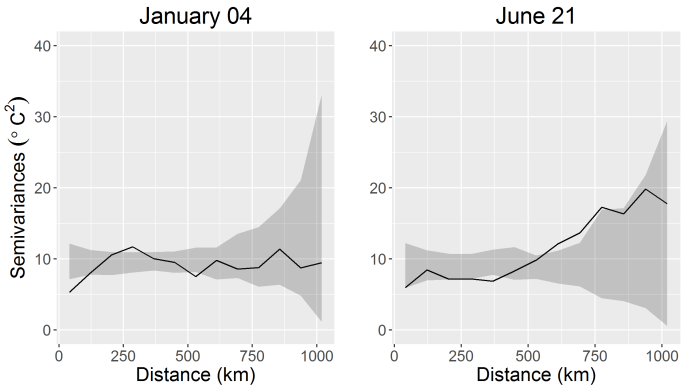


Figure 4.5: Binned empirical semivariograms with Monte Carlo envelopes in shaded area. These envelopes are obtained by permutations of the data across the sampling locations, and indicate regions of uncorrelated data.

Figure 4.6 illustrates how the latitude and longitude effects change over time. The longitude effect has a clearly increasing trend and this is possibly due to the Cascade mountains that extend from southern British Columbia through Washington and Oregon to Northern California. The Cascades block the westward movement of most of the cold, dense air that manages to reach eastern Washington and Oregon.

Our preliminary analysis indicates that for regions where topography changes significantly, simple polynomial trends commonly used in practice may introduce bias in the spatio–temporal residuals resulting in a semivariogram with a large squared bias term that can lead to a spurious finding of nonstationarity when none exists. Thus, our preliminary analysis points to the need for the improved estimation of the spatial mean model as reported in the sequel, one that accounts for extra features, notably elevation.

In particular, we recognize that our analysis needs to include spatio–temporal interactions as well as a longitude–elevation interaction. The latter

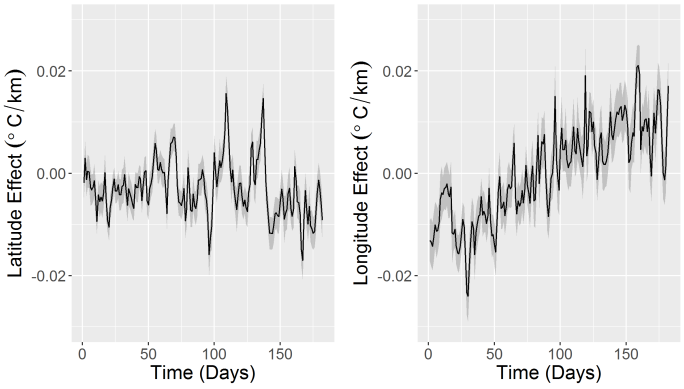


Figure 4.6: Latitude and longitude effects changing over time. The shaded area represents 95% confidence intervals for these effects.

is due to the effect of the proximity to the Pacific Ocean, which also takes into consideration the elevation effect due to the mountain ranges when moving eastward. Moreover, the longitude effect is assumed to depend on the elevation as well as how far north the station is located, and this effect must be allowed to vary over time. Similarly, the latitude effect may vary over time and it is dependent on how far east the weather station is located.

The above considerations lead to a spatio–temporal mean (trend) function that may be described as follows:

$$\mu_t(\mathbf{s}) \quad \equiv \quad f(\texttt{long*lat*month,long*elev}), \tag{4.3}$$

where f denotes a linear function, $\mathbf{s} = (\texttt{long}, \texttt{lat})$ are projected latitude and longitude coordinates, month indicates the month in study, and `elev` the elevation at $\mathbf{s}$. The $*$ notation is used to indicate that the mean function includes the individual, the two-way and, where applicable, the three-way interaction effects.

However, an alternative method suggests itself, one based on the use of historical temperature averages over the region to account for these complex interactions, as a representation of the climate in the Pacific Northwest. We describe this alternative in the following Subsection 4.3.3.

4.3.3 PRISM Climate Group Data

PRISM is a climate analysis system that uses point data, a digital elevation model (DEM), i.e. digital representations of cartographic information in a raster form, and other spatial data to generate gridded estimates of annual, monthly and event-based climatic parameters (Daly et al., 1994, 1997). It was developed primarily to interpolate climate elements in physiographically complex landscapes (Daly et al., 2008), and is particularly useful to identify short and long-term climate patterns.

The extrapolation of climate over high elevation ranges is often needed due to the lack of observations in mountainous regions. The use of PRISM data would then be ideal for complex regions with mountainous terrain such as the Pacific Northwest. In the literature, Daly et al. (1994, 1997, 2000) provide a description of the methodology behind PRISM. The main idea is that it calculates linear parameter–elevation relationships, allowing the slope to change locally with elevation. Observations nearer to the target elevation receive more weight than those further up or down slope. For temperature, the elevation of the top of the boundary layer is estimated by using the elevation of the lowest DEM pixels in the vicinity and adding a climatological inversion height to this elevation. (Daly et al., 1997)

The PRISM data we obtained corresponds to average values for temperature computed over a 30-year range (1981-2010), provided by the PRISM Climate Group, Oregon State University, and available online at `http://prism.oregonstate.edu` (last accessed on June 15, 2016). Our goal is to use these data as a representation of the climate in the Pacific Northwest. Having this information enables a comparison with our observations and an analysis of anomalies (i.e. differences between actual and expected values via PRISM), which would highlight what could not have been explained by the expected climate. This will also serve as a baseline comparison with the more complex mean function proposed in Section 4.3.2 that includes spatio–temporal interactions. Finally, in the sequel, PRISM data are used to construct a spatio–temporal trend model as an alternative to that suggested by the analysis reported in Section 4.3.2.

4.4 Bayesian Spatial Prediction

In this section, we present an empirical Bayesian spatial prediction (BSP) method built on the assumption that realizations of an underlying random field are obtained from measurements made at g gauged stations and that the goal is to obtain spatial predictions at the other u ungauged stations. Let $\mathbf{Y}_t \equiv (\mathbf{Y}_t^{(u)}, \mathbf{Y}_t^{(g)})$ denote a p-dimensional row vector ($p = u + g$), where $\mathbf{Y}_t^{(u)}$ and $\mathbf{Y}_t^{(g)}$ corresponds to the row vectors at the ungauged and gauged stations, respectively. The variables $\mathbf{Y}_t$ are assumed to be independent over time, or have passed a pre-filtering preliminary step, such that for $t = 1, \ldots, n$,

$$\mathbf{Y}_t | \mathbf{z}_t, \mathbf{B}, \mathbf{\Sigma} \quad \sim \quad \mathcal{N}_p(\mathbf{z}_t \mathbf{B}, \mathbf{\Sigma}), \tag{4.4}$$

where $\mathcal{N}$ denotes a multivariate normal distribution, with subscripts making the dimension explicit; the $\mathbf{z}_t$ is a k-dimensional row vector of covariates and $\mathbf{B}$ denotes a $(k \times p)$ matrix of regression coefficients.

As originally formulated in Le and Zidek (1992), in the BSP modelling, covariates were allowed to vary with time, but not space. Over the ensuing decade the BSP was extended in a variety of ways as summarized in Le and Zidek (2006). In particular, the response vector at each space–time point could be multivariate, thus enabling site specific random covariates with a Gaussian distribution to be incorporated in the BSP by first including them in the fitted multivariate joint distribution in Equation (8.44) and then conditioning on them to get the BSP. However, no way had been found to incorporate site specific nonrandom covariates.

However, such covariates are confronted in our analysis of temperature fields in complex regions, as the spatio-temporal mean function must include, say, topographic features as well as the oftentimes crucial spatio-temporal interactions. Thus, an extension was needed and the one that we developed, will now be presented.

Let $\mathbf{Y}$ be a $(n \times p)$ response matrix such that $\mathbf{Y} \equiv (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$, $\mathbf{Z}$ is a $(n \times k)$ design matrix and $\mathbf{B}$ a $(k \times p)$ matrix of regression coefficients.

Assume that

$$\begin{aligned} \mathbf{Y}|\mathbf{Z},\mathbf{B},\mathbf{\Sigma} &\sim \mathcal{MN}_{n\times p}(\mathbf{ZB},\mathbf{I},\mathbf{\Sigma}) && (4.5)\\ \mathbf{B}|\mathbf{B}_0,\mathbf{\Sigma},\mathbf{F} &\sim \mathcal{MN}_{k\times p}(\mathbf{B}_0,\mathbf{F}^{-1},\mathbf{\Sigma}) && (4.6)\\ \mathbf{\Sigma} &\sim \mathcal{W}_p^{-1}(\mathbf{\Xi},\delta), && (4.7) \end{aligned}$$

where $\mathbf{F}^{-1}$ is a positive $(k\times k)$ definite matrix, and $\mathbf{\Xi}$ a $(p\times p)$ hyperparameter matrix. Here, $\mathcal{MN}$ and $\mathcal{W}^{-1}$ denote the matrix normal and the inverted Wishart distributions, respectively, with subscripts making the dimensions explicit. We write

$$\mathbf{B}_0 = \begin{pmatrix} \beta_0^{(1)} & \dots & \beta_0^{(p)} \\ \beta_1^{(1)} & \dots & \beta_1^{(p)} \\ \vdots & & \vdots \\ \beta_{k-1}^{(1)} & \dots & \beta_{k-1}^{(p)} \end{pmatrix}, \tag{4.8}$$

where $\beta_0^{(j)} = \alpha + \sum_l \beta_{z_l} z_{l_j}$ includes the site–specific covariates at site j, denoted as z_{l_j}, $j = 1,\dots,p$ and for $i = 1,\dots,k$, $\beta_i^{(j)}$ denotes the coefficients of the non-site specific covariates. The first column of the Z matrix corresponds to a unit column vector, whereas the subsequent columns would contain the non–site specific covariates. Note that the method entails "burying" the site–specific covariates in the intercepts $\beta_0^{(j)}$. In practice, all of the regression parameters are first estimated via least squares, and ultimately plugged into the $\mathbf{B}_0$ matrix.

Denoting the matrices $\mathbf{\Sigma}_{gg}$ and $\mathbf{\Sigma}_{uu}$ as the covariance matrices of $\mathbf{Y}^{(g)}$ and $\mathbf{Y}^{(u)}$, respectively, and $\mathbf{\Sigma}_{ug}$ the cross-covariance, we can partition $\mathbf{\Sigma}$ and similarly the hyperparameter matrix $\mathbf{\Xi}$ as

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{uu} & \mathbf{\Sigma}_{ug} \\ \mathbf{\Sigma}_{gu} & \mathbf{\Sigma}_{gg} \end{pmatrix} \text{ and } \mathbf{\Xi} = \begin{pmatrix} \mathbf{\Xi}_{uu} & \mathbf{\Xi}_{ug} \\ \mathbf{\Xi}_{gu} & \mathbf{\Xi}_{gg} \end{pmatrix}. \tag{4.9}$$

For a fully (proper) Bayesian approach, extra hierarchy levels could be specified. Nonetheless, the BSP was developed from its inception to save

computational time by bypassing this approach. It was recognized that, in practice, the lack of prior knowledge would inevitably lead to a somewhat arbitrary choice of a convenience prior in this high-dimensional model. Thus, a preliminary empirical Bayes step is required for estimating $\mathbf{B}_0$ via a linear regression modelling approach, as suggested by the preliminary analysis in Section 4.3.2.

When performing spatial prediction, we use the result showed in Le and Zidek (1992) that the conditional distribution of $\mathbf{Y}_t^{(u)}$, where $t \in \{1, \ldots, n\}$ is given by

$$\mathbf{Y}_t^{(u)} | \mathbf{y}_t^{(g)}, \mathbf{Z}, \mathbf{B}_0 \sim t_u \left(\boldsymbol{\mu}^{(u)}, \frac{d}{\delta - u + 1} \boldsymbol{\Xi}_{u|g}, \delta - u + 1 \right), \tag{4.10}$$

where

$$\begin{aligned} \boldsymbol{\mu}^{(u)} &= \mathbf{z}_t \mathbf{B}_0^{(u)} + \boldsymbol{\Xi}_{ug} \boldsymbol{\Xi}_{gg}^{-1} (\mathbf{y}_t^{(g)} - \mathbf{z}_t \mathbf{B}_0^{(g)}) & (4.11) \\ d &= 1 + \mathbf{z}_t \mathbf{F}^{-1} \mathbf{z}_t^\top + (\mathbf{y}_t^{(g)} - \mathbf{z}_t \mathbf{B}_0^{(g)}) \boldsymbol{\Xi}_{gg}^{-1} (\mathbf{y}_t^{(g)} - \mathbf{z}_t \mathbf{B}_0^{(g)})^\top & (4.12) \\ \boldsymbol{\Xi}_{u|g} &= \boldsymbol{\Xi}_{uu} - \boldsymbol{\Xi}_{ug} \boldsymbol{\Xi}_{gg}^{-1} \boldsymbol{\Xi}_{gu}. & (4.13) \end{aligned}$$

Here $\mathbf{B}_0$ was partitioned as $\mathbf{B}_0 = (\mathbf{B}_0^{(u)}, \mathbf{B}_0^{(g)})$ according to the partition of $\mathbf{Y}_t$ (superscripts denoting the ungauged and gauged parts).

To finish model development, the covariance of the residual responses in Equation (8.44), $\boldsymbol{\Xi}_{u|g}$, must be specified. However, in practice and in our applications, these residuals will not have a second-order stationary distribution. Thus, we need to handle residuals with a non-stationary distribution. In this work, we adopt the Sampson-Guttorp (SG) warping method (Sampson and Guttorp, 1992), as described in Section 2.1.1.

4.5 Results

This section presents the results of applying the BSP method described in Section 4.4. This is implementable using the `EnviroStat v0.4-0 R` package (Le et al., 2014), including the extension proposed. For this purpose, we initially selected 64 stations at random for training, leaving the remainder

of the 97 stations for validation purposes, as illustrated in Figure 4.7.

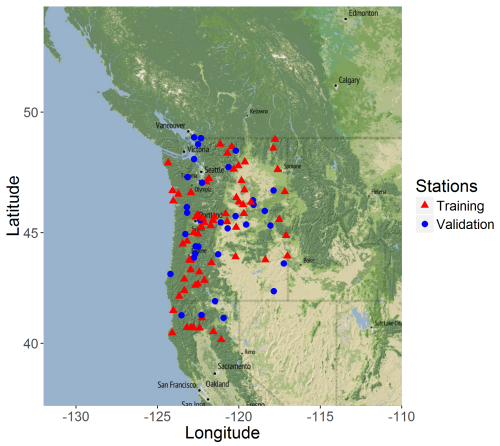


Figure 4.7: Locations of the stations selected for training and for validation purposes. Map created using `R` library `ggmap` with tiles by Stamen Design, under CC BY 3.0, and data by OpenStreetMap, under CC BY SA.

Work begins with an analysis of the spatio-temporal trend, as it is described in the subsequent Section 4.5.1, followed by an analysis of the spatial correlation in the residuals, after taking into account the mean trend in Section 4.5.2.

4.5.1 The Spatio–temporal Trend

For the training stations, Figure 4.8 illustrates the effect of projected latitude on temperatures considering different scenarios of longitude, elevation and time. Notice that moving north implies that the temperature in fact decreases in different rates, depending on your initial scenario. We refer to this as the RC-effect, which relates to a phenomenon where the effect of latitude and longitude change over time, that is, at certain times, widely separated sites might be strongly correlated. In a statistical model, this effect alerts to the need to include space-time interactions.

Figure 4.9 indicates that the Gaussian assumption is reasonably met, so no transformation is required.

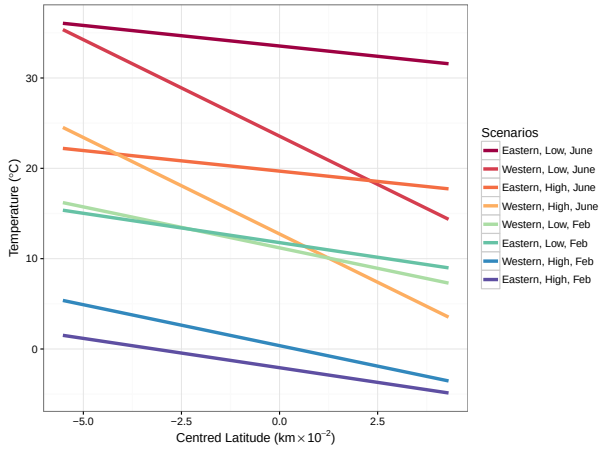


Figure 4.8: Investigating effect of projected latitude (centred) on temperatures considering different interaction scenarios for longitude (eastern, western), elevation (high, low) and time (February, June).

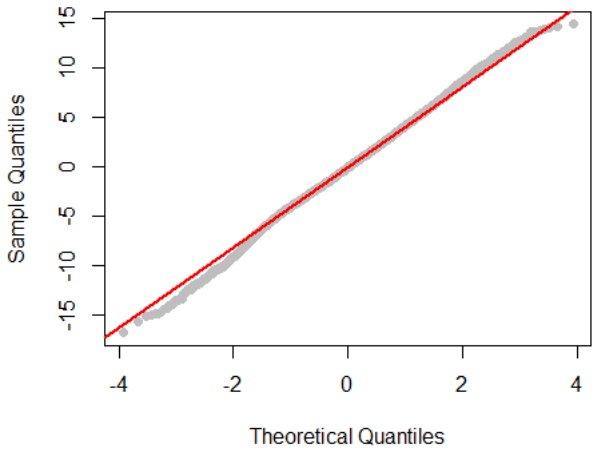


Figure 4.9: Normal quantile-quantile plot of the residual temperatures of a linear model with spatio-temporal mean function as in Equation 4.3.

The following Subsection 4.5.2 provides an analysis of the spatial correlation in the residuals, after taking into account this spatio-temporal trend.

4.5.2 Spatial Correlation in the Residuals

An important diagnostic in applying the SG method, supplied with the `EnviroStat v0.4-0 R` package (Le et al., 2014), is the biorthogonal grid seen in Figure 4.10. It represents the degree of contracting and expanding of the $\mathcal{G}$-space needed to attain an approximately stationary domain in $\mathcal{D}$-space through deformation. The solid lines indicate contraction and dashed lines, expansion. The expansions can be explained by the abrupt changes in the residual temperatures for nearby regions, due to diverse terrain. A contraction is seen in Eastern Washington, a basin located between the Cascade and Rocky Mountains.

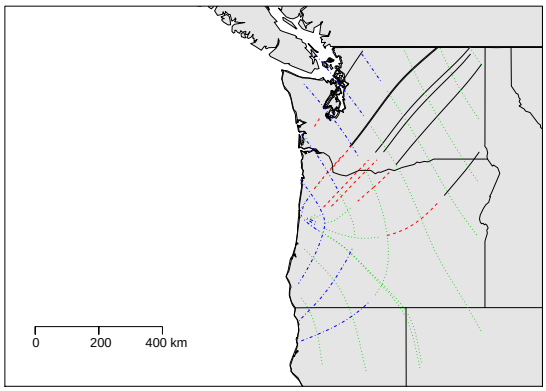


Figure 4.10: Biorthogonal grid for the thin-plate spline characterizing the deformation of the $\mathcal{G}$-space, using NCDC data set. Solid line indicates contraction and dashed lines indicate expansion.

Figure 4.11 illustrates the effect of different spline smoothing λ values in the deformed space. Without any smoothing ($\lambda = 0$), the $\mathcal{D}$-space is folded over on itself, implying that widely separated sites tend to be more correlated than sites located between them. To make the results more interpretable, we have chosen $\lambda = 5$, a value that keeps more of the gains from deformation seen in Figure 4.12, without folding the $\mathcal{G}$-space.

Figure 4.12 contains estimated dispersions after applying the SG ap-

proach (Sampson and Guttorp, 1992) in $\mathcal{G}$-space and $\mathcal{D}$-space. A more stationary fit is seen in the distorted space, since less variability is seen around the (stationary) variogram line.

We repeated the analysis using the PRISM data described in Section 4.3.3. This corresponds to average values for temperature computed over a 30-year range (1981-2010), and we use these data to represent the expected climate in the Pacific Northwest. Instead of estimating trend coefficients based on Equation 4.3, we analyze anomalies (i.e. differences between our observed values and those via PRISM). The goal is to validate our estimated trend by comparing improvements in prediction between these different analyses.

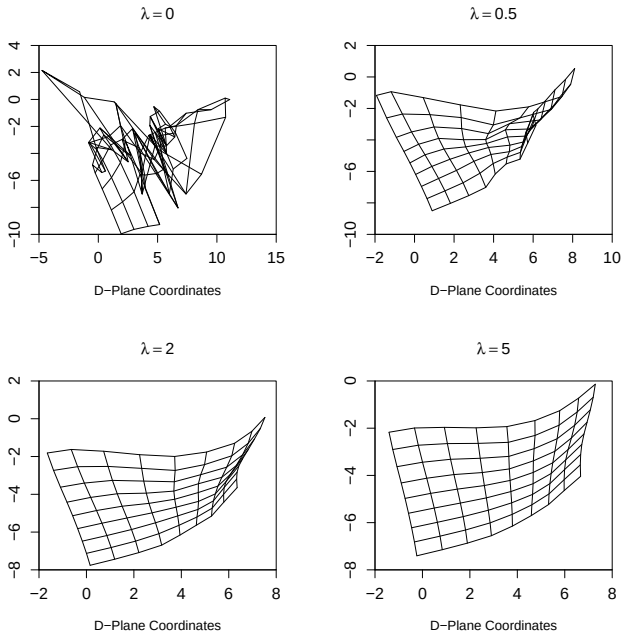


Figure 4.11: Deformation assuming different spline smoothing λ values. Note that when $\lambda = 0$, no smoothing is applied.

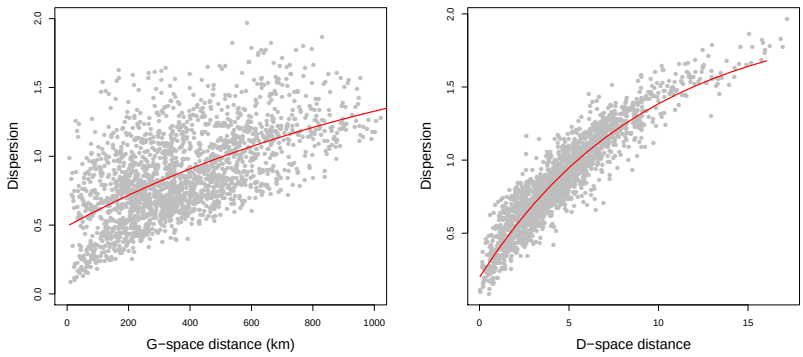


Figure 4.12: Estimated dispersions after SG approach in $\mathcal{G}$-space and in $\mathcal{D}$-space. The solid line represents a fitted exponential variogram.

In the following section, we assess and compare the spatial predictions made by our fitted spatio–temporal model with our two alternative approaches for modelling the spatio–temporal trend. We argue that PRISM captures the large-scale trend well, but it may not capture the effects of terrain at smaller scales.

4.5.3 Spatial Prediction

In this section we present our assessments of the prediction accuracy of our fitted hierarchical spatio–temporal model. For validation purposes, we compare the predicted values with the real values observed for the 33 left-out stations. Figure 4.13 contains a map of the mean squared prediction errors, averaged over time.

The prediction accuracy of the hierarchical spatio-temporal Bayesian model is also compared to ordinary kriging. For ordinary kriging, we used the `geoR v.1.7-4.1 R` package (Ribeiro Jr. and Diggle, 2001). The parameter estimates of the Exponential covariance function were obtained via maximum likelihood for the different time points. Figure 4.14 displays the mean squared prediction error for the ungauged stations and at different time points, respectively. From Table 4.1, notice that the coverage for our space-time interaction spatial mean is similar when analyzing PRISM anomalies,

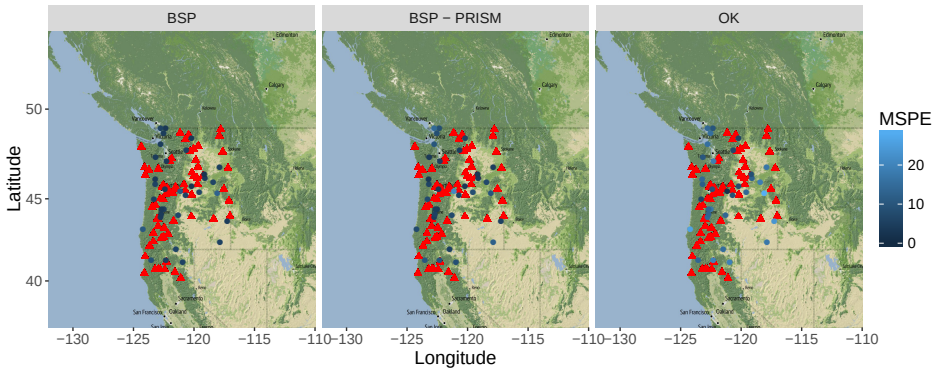


Figure 4.13: Map of mean squared prediction errors ($°C^2$), MSPE, averaged over time for the different methods considered: Bayesian spatial prediction (BSP), Bayesian spatial prediction with PRISM (BSP – PRISM), and ordinary kriging (OK). The red triangles represent the stations used for training purposes. Map created using `R` library `ggmap` with tiles by Stamen Design, under CC BY 3.0, and data by OpenStreetMap, under CC BY SA.

which serves as a way to characterize the strength of our general temperature mapping theory.

Table 4.1: Empirical coverage probabilities and prediction summaries for the different methods considered: Bayesian spatial prediction (BSP), Bayesian spatial prediction with PRISM (BSP – PRISM), and ordinary kriging. The overall MSPE refers to the mean squared prediction errors ($°C^2$) averaged over space and time.

	BSP	BSP – PRISM	Ordinary kriging
Empirical coverage probabilities of 95% CI	0.918	0.921	0.529
Overall MSPE	5.396	7.000	14.032
(std.error)	(2.362)	(3.733)	(5.823)

In addition, Figure 4.14 shows that the mean squared prediction errors

across ungauged stations and across time are, on average, smaller for the BSP method introduced considering the spatio-temporal interactions in the mean function as in Equation 4.3. The reason for this could be due to the fact that PRISM may not be capturing the effects of terrain at smaller scales.

Another disadvantage is that the PRISM data are currently not available at locations outside of the United States. Thus, we advocate that for regions with complex terrain like the Pacific Northwest, a thorough exploratory analysis is crucial to better understand the local changes in trend.

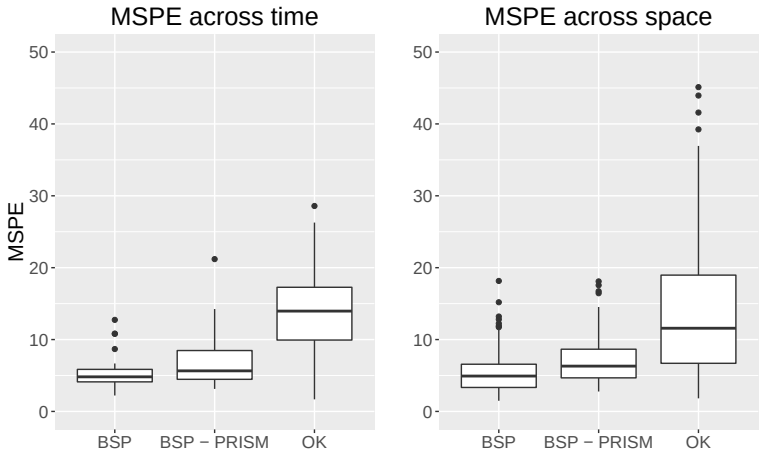


Figure 4.14: Mean squared prediction errors ($^{\circ}C^2$) averaged across ungauged stations and across time for the different methods considered: Bayesian spatial prediction (BSP), Bayesian spatial prediction with PRISM (BSP - PRISM), and ordinary kriging (OK).

4.6 Concluding Remarks

This chapter focused on the modelling temperature fields in the Pacific Northwest, where rapid changes in temperature and localized weather are common particularly due to the complex terrain.

We introduced a flexible stochastic spatio-temporal model for daily tem-

peratures in the Pacific Northwest that handles nonstationarity. We also stressed the need for spatio-temporal interactions to understand the temperature trends. We believe that global climate models may fail to explain interesting smaller-scale trends, especially in regions with a complex terrain like the Pacific Northwest.

In addition, we introduced two comparable strategies for spatial prediction in regions with a complex terrain. The first is an extension of the Bayesian spatial prediction proposed by Le and Zidek (1992). We extended this method to take into account spatio-temporal interaction features in the mean to capture the localized changes in trend. The second is based on tackling the anomalies of the expected climate in the Pacific Northwest, based on the average values of temperature computed over a 30-year range (1981-2010), provided by PRISM Climate Group. PRISM data, however, are currently not available at locations outside of the United States.

This work conclusively shows how appropriately modelling the spatio-temporal mean field can help resolve these complex patterns for nonstationarity and improve spatial prediction in contrast to using simpler mean structures. This can be seen by larger MSPEs observed for the BSP-PRISM, where anomalies of expected weather were analyzed, instead of investigating observed changes in temperature across the Pacific Northwest. Moreover, we would like to emphasize the need to account for nonstationarity in this region, as demonstrated by the underperformance of the more traditional kriging methodology based on stationary models.

Our analysis also discovered abrupt changes in the observed temperatures for nearby regions due to diverse terrain in a great part of the western region, and less variable weather conditions in Eastern Washington, a basin located between the Cascade and Rocky Mountains.

Chapter 5

Ensemble Modelling

5.1 Motivation

The use of computer models has become increasingly common in environmental science applications. These computer models are used to simulate physical phenomena in order to better understand complex physical systems. From a statistical perspective, the focus is often on processing information brought by their outputs and gathering useful insights about the underlying system. In practice, environmental agencies often depend on these kinds of information for regulatory purposes.

It should be noted, however, that these outputs come from deterministic models, and hence there are no indications of uncertainty associated with them. Kennedy and O'Hagan (2001) dealt with uncertainty analysis and introduced a Bayesian calibration technique by incorporating information from both computer model outputs and monitoring data.

One of the recent challenges in spatial statistics applications is in fusing information from multiple sources that might have been measured at different spatial scales. This problem is often seen as data fusion and it is also referred to as the change of support problem. Although the issue of handling mismatched spatial scales is a well established problem, recent advances in remote sensing highlights the need for suitable statistical methods to address it (Nguyen et al., 2012). A comprehensive review of methods for dealing with mismatched spatial data can be found in Gotway and Young (2002) and Gelfand (2010). In this chapter, we build upon Fuentes and Raftery (2005), where we handle different spatial scales by linking them through an underlying "true" process at the micro-scale.

Notably, in weather studies, data often come from monitoring stations,

as it was the case in Chapter 4, but supplemented by the inclusion of the computer model outputs in the modelling. In this chapter, we will explore these ideas for combining multiple computer models for a temperature data set in the Pacific Northwest. As noted in Nychka and Anderson (2010), numerical weather prediction is one of the most successful applications of the data fusion problem, where a large set of observations are combined with physical models describing the atmosphere evolution and ultimately producing high resolution weather forecasts.

Other strategies that do not assume an underlying "true" process include downscaler models (Berrocal et al., 2010a,b, 2012) and a two-stage regression approach (Guillas et al., 2006, 2008; Zidek et al., 2012). The downscaling strategy handles the station and model outputs observations at mismatched scales via a regression model with spatially varying coefficients. The two-stage regression approach as in Guillas et al. (2006, 2008) is based on first regressing the station data on the model outputs and then regressing the estimated residuals of the first step on indicators of time and other temporal components. The work in Zidek et al. (2012) provides an extension based on an ad-hoc method to allow spatial interpolation of the coefficients of the linear regression.

5.1.1 Contributions

Ensemble modelling is hereby referred to as a statistical post-processing technique based on combining multiple computer models outputs in a statistical model with the goal of obtaining probabilistic forecasts.

In Section 5.2, we provide a description of the Bayesian Ensemble Melding model (BEM) methodology introduced by Liu (2007) following Fuentes and Raftery (2005). The main idea lies in linking processes on mismatched scales through an underlying "true" process. One of the main disadvantages of these methodologies is the computational burden faced while performing inference, as noted in many applications (Swall and Davis, 2006; Smith and Cowles, 2007; Foley and Fuentes, 2008).

Moreover, simple MCMC strategies are known to be infeasible when

handling large spatial data sets. In this chapter, our main objective is to introduce a scalable inference methodology alternative for the BEM using integrated nested Laplace approximations described in Section 3.2. We follow Lindgren et al. (2011) and take advantage of a Markov representation of the Matérn covariance family, connecting ideas from Gaussian Markov random fields (GMRFs) and stochastic partial differential equations (SPDEs) in a continuous space.

Since the BEM is essentially a spatial model, our ultimate goal is to provide some background to Chapter 6, where we build upon the ability of the BEM to accommodate time and describe a dynamic strategy with the objective of performing forecasting. In that chapter, we will take advantage of the computational gains of the INLA methodology for performing inference for the BEM. On the other hand, McMillan et al. (2010) proposed a spatio-temporal extension through a specification of the underlying "true" process at a grid cell scale. They made use of MCMC methods for performing inference and due to the large number of grid cells, the computational burden was still an issue.

5.2 The Bayesian Ensemble Melding Model

The Bayesian Ensemble Melding (BEM) model described in Liu (2007) can be viewed as an extension of the Bayesian model proposed by Fuentes and Raftery (2005), by allowing the combination of the observed measurements with outputs from an ensemble of deterministic models.

Similarly to Fuentes and Raftery (2005), the BEM model is able to link processes on mismatched scales through an underlying "true" process $\{Z(\mathbf{s}) : \mathbf{s} \in \mathbb{R}^d\}$, where d is the dimension of the domain, and through the consideration of the conceptual processes $\{\tilde{Z}_j(\mathbf{s}) : \mathbf{s} \in \mathbb{R}^d\}$, the deterministic model output processes, $j = 1, \ldots, p$.

The different spatial scales are dealt with by linking the underlying "true" process in a micro-scale with the model outputs, which may be averages at a grid-scale resolution. A similar idea was used in Wikle and Berliner (2005), where they describe it as a conditional change of support solution.

Wikle and Berliner (2005) handle the mismatched scales by conditioning a true unobserved spatially continuous process on an areal average of the process at some support in which there is interest in performing inference.

For the BEM model, at a given location $\mathbf{s} \in \mathbb{R}^d$, the measurement process is modeled as

$$\hat{Z}(\mathbf{s}) = Z(\mathbf{s}) + e(\mathbf{s}), \tag{5.1}$$

where the measurement error process at location $\mathbf{s}$ is assumed to be $e(\mathbf{s}) \sim N(0, \sigma_e^2)$, and independent of $Z(\mathbf{s})$. We represent the realizations of the measurement process at all locations as the vector $\hat{\mathbf{Z}}$.

In addition, for each $j = 1, \ldots, p$, the j-th output process from the deterministic models $\{\tilde{Z}_j(\mathbf{s}) : \mathbf{s} \in \mathbb{R}^d\}$ is modeled as:

$$\tilde{Z}_j(\mathbf{s}) = a_j(\mathbf{s}) + b_j(\mathbf{s})Z(\mathbf{s}) + \delta_j(\mathbf{s}), \tag{5.2}$$

where for each $j = 1, \ldots, p$, the error term is $\delta_j(\mathbf{s}) \sim N(0, \sigma_{\delta,j}^2)$. The parameter functions a_j and b_j measure the additive and multiplicative calibration parameters for the j-th deterministic model. The processes $\delta_j(\cdot)$ are assumed independent of each other as well as independent of $e(\mathbf{s})$, the measurement error process.

Since the deterministic model outputs are generally measured in sub-regions $B_1, \ldots, B_m$ (blocks) of the study domain, for each deterministic model $j = i, \ldots, p$, and each sub-region $i = 1, \ldots, m$,

$$\tilde{Z}_j(B_i) = \frac{1}{|B_i|}\left(\int_{B_i} a_j(\mathbf{s})d\mathbf{s} + b_j \int_{B_i} Z(\mathbf{s})d\mathbf{s} + \int_{B_i} \delta_j(\mathbf{s})d\mathbf{s}\right), \tag{5.3}$$

where $|B_i|$ denotes the area of the sub-regions of the study domain for $i = 1, \ldots, m$.

One way of approximating $\int_{B_i} Z(\mathbf{s})d\mathbf{s}$ is by Monte Carlo integration, after obtaining a sample of locations over the $B_1, \ldots, B_m$. To this end, suppose that a random sample of L points are obtained in each of sub-region

$B_1, \ldots, B_m$ and thus for $i = 1, \ldots, m$

$$Z(B_i) = \int_{B_i} Z(\mathbf{s}) d\mathbf{s} \tag{5.4}$$

$$\approx \frac{1}{L} \sum_{k=1}^{L} Z(\mathbf{s}_{k,B_i}) d\mathbf{s}. \tag{5.5}$$

Originally, Liu (2007) assumed no overlap between these sampling locations and the monitoring sites within the sub-regions $B_1, \ldots, B_m$. We represent the realizations for the j-th deterministic model output at all mL sampling locations as the vector $\tilde{\mathbf{Z}}_j$. In the case where the deterministic models were previously interpolated at the monitoring locations, $\tilde{\mathbf{Z}}_j$ has the same dimension as $\hat{\mathbf{Z}}$.

Finally, in order to link the above processes, the "true" underlying process is modeled as

$$Z(\mathbf{s}) = \mu(\mathbf{s}) + \epsilon(\mathbf{s}), \tag{5.6}$$

where $\mu(\mathbf{s})$ is a deterministic mean structure, usually a polynomial function of $\mathbf{s}$, representing large-scale variation. The mean parameters are denoted as $\boldsymbol{\beta}$. The errors $\epsilon(\mathbf{s})$ are assumed to be correlated with zero mean, and variance denoted as σ^2. Their correlation could be described by a parametric correlation functions for stationary processes or a non-stationary structure in a more general case. These correlation parameters are denoted as $\boldsymbol{\theta}$.

Realizations of the "true" underlying process at all locations are represented in the vector $\mathbf{Z}$. Key to linking the measurement and deterministic processes is that not only these realizations of the "true" underlying process need be observed at the n monitoring stations, but also at the sampling locations where the deterministic model processes are observed. In the case where the p deterministic model processes are observed at L locations within each sub-region $B_1, \ldots, B_m$, the vector $\mathbf{Z}$ has dimension $(n + pmL) \times 1$.

Due to the mismatched dimensions, the linking matrices $\mathbf{A}_0$ and $\mathbf{A}_j$, for

each $j = 1, \ldots, p$ are introduced such that

$$\dim(\mathbf{A}_0\mathbf{Z}) = \dim(\hat{\mathbf{Z}}) = (n \times 1) \tag{5.7}$$

$$\dim(\mathbf{A}_j\mathbf{Z}) = \dim(\tilde{\mathbf{Z}}_j) = (m \times 1). \tag{5.8}$$

First, consider the following auxiliary block matrix, with as many blocks as there are deterministic models

$$\mathbf{E}_j = \left[\mathbf{0}^{(1)} \ldots \mathbf{0}^{(j-1)}\mathbf{L}^{(j)}\mathbf{0}^{(j+1)} \ldots \mathbf{0}^{(p)}\right], \tag{5.9}$$

and where each of the blocks corresponds to a $(m \times mL)$ and the j-th block is given by the following

$$\mathbf{L}^{(j)} = \begin{pmatrix} \frac{1}{L} & \ldots & \frac{1}{L} & \ldots & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \ldots & 0 & \ldots & \frac{1}{L} & \ldots & \frac{1}{L} \end{pmatrix}. \tag{5.10}$$

The superscripts are used to differentiate the different blocks in the $\mathbf{E}$ matrix. There are a total of p blocks, thus $\mathbf{E}_j$ has dimensions $(m \times pmL)$. Then we can finally define the linking matrices as

$$\mathbf{A}_0 = [\mathbf{I}_n|\mathbf{0}_{(n\times pmL)}] \tag{5.11}$$

$$\mathbf{A}_j = \left[\mathbf{0}_{(m\times n)}|\mathbf{E}_j\right], \tag{5.12}$$

where $\mathbf{I}_n$ denotes a $(n \times n)$ identity matrix, and $\mathbf{0}_{(i\times j)}$ a $(i \times j)$ zero matrix. Hence, $\mathbf{A}_0$ has dimensions $(n \times (n + pmL))$ and for each $j = 1, \ldots, p$, $\mathbf{A}_j$ has dimensions $(m \times (n + pmL))$.

In the following Section 5.3, we discuss the inference for the BEM model.

5.3 Inference for the BEM

Let $\boldsymbol{\Psi} = (\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2, a_1, \ldots, a_p, b_1, \ldots, b_p, \sigma^2_{\delta_1}, \ldots, \sigma^2_{\delta_p}, \sigma^2_e)$ denote all model parameters. Summarizing the BEM, note that

$$\begin{aligned}
\hat{\mathbf{Z}}|\mathbf{Z}, \boldsymbol{\Psi} &\sim \mathcal{N}(\mathbf{A}_0\mathbf{Z}, \sigma^2_e\mathbf{I}_n) \\
\tilde{\mathbf{Z}}_j|\mathbf{Z}, \boldsymbol{\Psi} &\sim \mathcal{N}(a_j\mathbf{1}_m + b_j\mathbf{A}_j\mathbf{Z}, \sigma^2_{\delta_j}\mathbf{I}_m) \\
\mathbf{Z}|\boldsymbol{\Psi} &\sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),
\end{aligned}$$

for $j = 1, \ldots, p$, where $\mathbf{1}_m$ denote a unit vector of size m. For model completeness, prior assumptions must be made for $\boldsymbol{\Psi}$.

Here, we consider calibration parameters a_j, b_j for $j = 1, \ldots, p$ are constant in space. For particular applications, it might be of relevance to let these parameters vary across space.

The posterior distribution of $\boldsymbol{\Psi}$ and the "true" spatial underlying process is given by

$$p(\boldsymbol{\Psi}, \mathbf{Z}|\hat{\mathbf{Z}}, \tilde{\mathbf{Z}}_{1:p}) \propto p(\hat{\mathbf{Z}}|\mathbf{Z}, \sigma^2_e) \prod_{j=1}^{p} \left[p(\tilde{\mathbf{Z}}_j|\mathbf{Z}, a_j, b_j, \sigma^2_{\delta_j})\right] p(\mathbf{Z}|\boldsymbol{\beta}, \boldsymbol{\theta}) p(\boldsymbol{\Psi}) \tag{5.13}$$

$$\begin{aligned}
&\propto p(\boldsymbol{\Psi})|(\sigma^2_e\mathbf{I}_n)|^{-\frac{1}{2}}|\Sigma|^{-\frac{1}{2}}[\prod_{j=1}^{p}|(\sigma^2_{\delta_j}\mathbf{I}_m)|^{-\frac{1}{2}}] \qquad (5.14) \\
&\exp\left\{-\frac{1}{2}\left[(\hat{\mathbf{Z}} - \mathbf{A}_0\mathbf{Z})^\top(\sigma^2_e\mathbf{I}_n)^{-1}(\hat{\mathbf{Z}} - \mathbf{A}_0\mathbf{Z})\right.\right. \\
&+(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta})^\top\Sigma^{-1}(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}) \\
&\left.\left. + \sum_{j=1}^{p}(\tilde{\mathbf{Z}}_j - \mathbf{1}a_j - b_j\mathbf{A}_j\mathbf{Z})^\top(\sigma^2_{\delta_j}\mathbf{I}_m)^{-1}(\tilde{\mathbf{Z}}_j - \mathbf{1}a_j - b_j\mathbf{A}_j\mathbf{Z})\right]\right\}
\end{aligned}$$

where $\hat{\mathbf{Z}} = (\hat{Z}(\mathbf{s}_1), \ldots, \hat{Z}(\mathbf{s}_n))^\top$, and $\tilde{\mathbf{Z}}_{1:p} = (\tilde{\mathbf{Z}}_1^\top, \ldots, \tilde{\mathbf{Z}}_p^\top)^\intercal$, where for each $j \in 1, \ldots, p$, $\tilde{\mathbf{Z}}_j = (\tilde{Z}_j(B_1), \ldots, \tilde{Z}_j(B_m))^\top$.

Liu et al. (2011) obtained samples from this posterior distribution by using MCMC methods. In fact, this has been the most common way of performing inference for such point-referenced spatial models. One of the main drawbacks of this approach is the computational burden associated with expensive matrix computations in each MCMC iteration, and the slow

mixing of the chains.

In Section 5.4, we describe an application of the BEM model based on INLA method for performing approximate Bayesian inference, as discussed in Section 3.2. In particular, in Subsection 5.3.1, we follow up on the work of Lindgren et al. (2011) and describe how to represent the BEM model in such framework.

5.3.1 A Stochastic-Partial Differential Equation Model Alternative

One of the main disadvantages of the BEM model is the computational burden associated with factorizing dense covariance structures for estimation and spatial prediction purposes. For large data sets, their use is impractical. Lindgren et al. (2011) proposed an alternative to this problem by taking advantage of a Markov representation of the Matérn covariance family, connecting ideas from between Gaussian Markov random fields (GMRFs) and stochastic partial differential equations (SPDEs) in a continuous space. Examples of applications using the `INLA-SPDE` approach for geostatistical models can be found in Simpson et al. (2012b,a) and Cameletti et al. (2013).

In this section, we provide an overview of Lindgren et al. (2011), and the discussion provided in Simpson et al. (2012b), later connecting the ideas into the BEM model. The key step entails replacing the dense covariance structure of a Gaussian field by a GMRF, thus allowing users to take advantage of sparse precision matrices. Ultimately, the INLA methodology (Rue et al., 2009) described in Chapter 3 could then be used for the purpose of inference. The `INLA-SPDE` approach can be implemented using the `INLA R` package, available for download at r-inla.org (last accessed on June 15, 2016).

Instead of the usual definition through the covariance function, Lindgren et al. (2011) suggest representing a Gaussian random field η with Matérn

covariance function as a solution to the SPDE

$$(\kappa^2 - \Delta)^{\frac{\alpha}{2}}\eta(\mathbf{s}) = \mathcal{W}(\mathbf{s}), \qquad (5.15)$$
$$\alpha = \nu + d/2,\ \kappa > 0,\ \nu > 0,$$

where $\Delta = \sum_{i=1}^{d} \frac{\partial^2}{\partial \eta_i^2}$ is the Laplacian operator and $\mathcal{W}$ a spatial Gaussian white noise with unit variance.

This is due to the fact that Gaussian random fields with Matérn covariance function are stationary solutions to (5.15) for any $\alpha \geq d/2$ (Whittle, 1954, 1963). When α is an integer, the Matérn fields are Markovian (Lindgren et al., 2011). The Matérn covariances are parametrized as

$$\text{Cov}(\mathbf{s}, \mathbf{s}+\mathbf{h}) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(\kappa||\mathbf{h}||)^{\nu} K_{\nu}(\kappa||\mathbf{h}||). \qquad (5.16)$$

Note that the two representations (5.15) and (5.16) are related, in such way that the Matérn smoothness is $\nu = \alpha - d/2$ and the marginal variance σ^2 is given by

$$\sigma^2 = \frac{\Gamma(\nu)}{\Gamma(\alpha)(4\pi)^{d/2}\kappa^{2\nu}}. \qquad (5.17)$$

Moreover, an empirically derived spatial range can be obtained from $\rho = \sqrt{\frac{8}{\kappa}}$, where the spatial correlation decays to approximately 13%.

For a finite set of suitable functions $\{\varphi_j(\mathbf{s}), j = 1, \ldots, N\}$, a solution of (5.15) satisfies

$$\int \varphi_j(\mathbf{s})(\kappa^2 - \Delta)^{\frac{\alpha}{2}}\eta(\mathbf{s})d\mathbf{s} = \int \varphi_j(\mathbf{s})\mathcal{W}(d\mathbf{s}). \qquad (5.18)$$

For instance, consider the case where $\alpha = 2$ on a two-dimensional domain $\Omega \subset \mathbb{R}^2$. Using Green's first identity (A.1), and assuming a zero normal derivative of $\eta(\mathbf{s})$ on the boundary of Ω, we can then rewrite (5.18) as

$$\int_{\Omega} [\kappa^2 \varphi_j(\mathbf{s})\eta(\mathbf{s}) + \nabla\varphi_j(\mathbf{s}) \cdot \nabla\eta(\mathbf{s})]d\mathbf{s} \quad = \quad \int_{\Omega} \varphi_j(\mathbf{s})\mathcal{W}(d\mathbf{s}). \qquad (5.19)$$

Since the idea is based on using GMRFs to approximate a Gaussian random field, Lindgren et al. (2011) then suggests that the Gaussian random field be approximated by a finite element representation of the solution of the SPDE based on a triangulation of the spatial domain. In `R-INLA`, a Delauney triangulation (DT), which maximizes the minimum interior triangle angle is used. Initial vertices are placed at the locations where observations are available. Then, additional vertices are added to cover the spatial domain, and finally yielding an irregular grid representation of the original (continuous) Gaussian process. The finite element representation of the solution of the SPDE in (5.15) is as follows

$$\eta(\mathbf{s}) = \sum_{k=1}^{N} \psi_k(\mathbf{s}) w_k, \tag{5.20}$$

for Gaussian weights $\{w_k\}$ and some basis functions $\{\psi_k\}$ which are piecewise linear in each triangle, defined such that ψ_k is 1 at vertex k and 0 at the other vertices.

In the case where $\alpha = 2$, we could obtain a set of N equations to solve by substituting (5.20) into (5.19)

$$\sum_{k=1}^{N} \left(\int_{\Omega} [\kappa^2 \varphi_j(\mathbf{s}) \psi_k(\mathbf{s}) + \nabla \varphi_j(\mathbf{s}) \cdot \nabla \psi_k(\mathbf{s})] d\mathbf{s} \right) w_k = \int_{\Omega} \varphi_j(\mathbf{s}) \mathcal{W}(d\mathbf{s}), \tag{5.21}$$

for $j = 1, \ldots, N$. Assuming that the test functions are the same as the basis functions, i.e. $\varphi_j(\mathbf{s}) = \psi_k(\mathbf{s})$, note that the right-hand size of (5.21) becomes $\int_{\Omega} \psi_j(\mathbf{s}) \mathcal{W}(d\mathbf{s})$, yielding a zero-mean normal distribution with covariances $\int_{\Omega} \psi_k(\mathbf{s}) \psi_j(\mathbf{s}) d\mathbf{s}$. Hence (5.21) can be rewritten as

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{\kappa^2} \mathbf{C}^{-1} \mathbf{K}_{\kappa^2}), \tag{5.22}$$

where $\mathbf{K}_{\kappa^2} = \kappa^2 \mathbf{C} + \mathbf{G}$. The above notation refers to a normal distribution with mean and precision given by the above. The matrices $\mathbf{K}_{\kappa^2}$, $\mathbf{C}$ and $\mathbf{G}$

have entries given by

$$
\begin{aligned}
(\mathbf{K}_{\kappa^2})_{ij} &= \kappa^2 \mathbf{C}_{ij} + \mathbf{G}_{ij} && (5.23)\\
\mathbf{C}_{ij} &= \int_{\Omega} \psi_i(\mathbf{s})\psi_j(\mathbf{s}) d\mathbf{s} && (5.24)\\
\mathbf{G}_{ij} &= \int_{\Omega} \nabla\varphi_j(\mathbf{s}) \cdot \nabla\psi_i(\mathbf{s}) d\mathbf{s}. && (5.25)
\end{aligned}
$$

Since $\mathbf{C}$ is dense, sparsity is imposed by replacing it with a diagonal matrix $\tilde{\mathbf{C}}$ with elements given by $\int_{\Omega} \psi_i(\mathbf{s}) d\mathbf{s}$, for $i = 1, \ldots, N$. Therefore $\mathbf{w}$ yields an approximate GMRF from the continuous Matérn field, which has a sparse precision matrix and is more computationally efficient than the typical dense Matérn one. In our applications, we assume $\alpha = 2$, hence our focus here in describing this case. For α equal to other integers, check Lindgren et al. (2011).

5.3.2 Spatial Prediction

Typical spatial prediction techniques via kriging normally involve estimating the parameters of the underlying covariance structure. Then these parameters are assumed known for use in spatial prediction. In this section, we describe a Bayesian approach to spatial prediction of the BEM which takes into account the uncertainty about parameters on these predictions.

Conditionally on $\boldsymbol{\Psi}$, the joint distribution of $\hat{\mathbf{Z}}$ and $\tilde{\mathbf{Z}}_{1:p}$ is multivariate normal, as follows:

$$
\begin{pmatrix} \hat{\mathbf{Z}} \\ \tilde{\mathbf{Z}}_1 \\ \tilde{\mathbf{Z}}_2 \\ \vdots \\ \tilde{\mathbf{Z}}_p \end{pmatrix} \sim \mathcal{N} \left\{ \begin{pmatrix} \hat{\boldsymbol{\mu}} \\ a_1 + b_1\tilde{\boldsymbol{\mu}} \\ a_2 + b_2\tilde{\boldsymbol{\mu}} \\ \vdots \\ a_j + b_j\tilde{\boldsymbol{\mu}} \end{pmatrix}, \begin{bmatrix} \Sigma_{\hat{\mathbf{Z}}} & \Sigma_{\hat{\mathbf{Z}},\tilde{\mathbf{Z}}_1} & \Sigma_{\hat{\mathbf{Z}},\tilde{\mathbf{Z}}_2} & \cdots & \Sigma_{\hat{\mathbf{Z}},\tilde{\mathbf{Z}}_p} \\ & \Sigma_{\tilde{\mathbf{Z}}_1} & \Sigma_{\tilde{\mathbf{Z}}_1,\tilde{\mathbf{Z}}_2} & \cdots & \Sigma_{\tilde{\mathbf{Z}}_1,\tilde{\mathbf{Z}}_p} \\ & & & \ddots & \vdots \\ & & \ddots & & \Sigma_{\tilde{\mathbf{Z}}_{p-1},\tilde{\mathbf{Z}}_p} \\ & & & & \Sigma_{\tilde{\mathbf{Z}}_p} \end{bmatrix} \right\} \tag{5.26}
$$

where $\hat{\boldsymbol{\mu}} = (\mu(\mathbf{s}_1), \ldots, \mu(\mathbf{s}_n))^{\intercal}$ and $\tilde{\boldsymbol{\mu}} = (\int_{B_1} \mu(\mathbf{s}) d\mathbf{s}, \ldots, \int_{B_m} \mu(\mathbf{s}) d\mathbf{s})^{\intercal}$. In the case that outputs have been previously interpolated to the observed

locations, then $\tilde{\boldsymbol{\mu}}$ simplifies to $\tilde{\boldsymbol{\mu}} = (\mu(\mathbf{s}_1), \ldots, \mu(\mathbf{s}_n))^\intercal$.

Regarding the covariance structure, $\Sigma_{\hat{\mathbf{Z}}}$ denotes the covariance matrix for $\hat{\mathbf{Z}}$, and for each $j = 1, \ldots, p$, $\Sigma_{\tilde{\mathbf{Z}}_j}$ denotes the covariance matrix for $\tilde{\mathbf{Z}}_j$. The matrix $\Sigma_{\hat{\mathbf{Z}},\tilde{\mathbf{Z}}_j}$ denotes the cross-covariance between the measurements $\hat{\mathbf{Z}}$ and the j-th output $\tilde{\mathbf{Z}}_j$, while $\Sigma_{\tilde{\mathbf{Z}}_i,\tilde{\mathbf{Z}}_j}$, $i \neq j$, denotes the cross-covariance between the i-th and j-th members of the ensemble. For $j = 1, \ldots, p$, and $i \neq j$, note that

$$\begin{aligned}
\Sigma_{\hat{\mathbf{Z}}} &= \mathbf{A}_0 \boldsymbol{\Sigma} \mathbf{A}_0^\top + \sigma_e^2 \mathbf{I}_n && (5.27)\\
\Sigma_{\tilde{\mathbf{Z}}_j} &= b_j^2 \mathbf{A}_j \boldsymbol{\Sigma} A_j^\top + \sigma_{\delta_j}^2 \mathbf{I}_m && (5.28)\\
\Sigma_{\hat{\mathbf{Z}},\tilde{\mathbf{Z}}_j} &= b_j \mathbf{A}_0 \boldsymbol{\Sigma} A_j^\top && (5.29)\\
\Sigma_{\tilde{\mathbf{Z}}_i,\tilde{\mathbf{Z}}_j} &= b_i b_j \mathbf{A}_j \boldsymbol{\Sigma} \mathbf{A}_j^\top, && (5.30)
\end{aligned}$$

where the $\mathbf{A}_0$ and $\mathbf{A}_j$ matrices have been described in (5.11) and (5.12).

In the case that the deterministic models have been previously interpolated to the measurement locations, these matrices may not need to be defined. In the context of prediction, however, we often predict temperature measurements conditionally on the model outputs. In practice, this means that over space, there are more sites where model outputs are available than temperature measurements. In such cases, it might be useful to define a matrix $\mathbf{A}_0$ responsible for extracting the part of the "true" underlying field that relates to the locations where temperature measurements are available.

We denote $\check{\boldsymbol{\mu}}$ and $\check{\boldsymbol{\Sigma}}$ as the mean vector and covariance matrix of the joint distribution of $\check{\mathbf{Z}} = (\hat{\mathbf{Z}}^\top, \tilde{\mathbf{Z}}_1^\top, \ldots, \tilde{\mathbf{Z}}_p^\top)^\top$ conditionally on $\boldsymbol{\Psi}$, as described in (5.26). For the purposes of spatial prediction, the goal is to predict $\hat{\mathbf{Z}}$ at a given set of m new locations $\mathbf{x}^*$. This entails describing the predictive distribution of $\hat{\mathbf{Z}}(\mathbf{x}^*)$. Recall from the model definition (5.6), conditionally on $\boldsymbol{\Psi}$, that the distribution of $\hat{\mathbf{Z}}(\mathbf{x}^*)$ is normal with mean $\hat{\boldsymbol{\mu}}(\mathbf{x}^*)$ and covariance matrix $\mathbf{A}_0 \boldsymbol{\Sigma} \mathbf{A}_0^\top + \sigma_e^2 \mathbf{I}_m$.

Now denote the cross-covariance between of $\hat{\mathbf{Z}}(\mathbf{x}^*)$ and $\hat{\mathbf{Z}}, \tilde{\mathbf{Z}}_{1:p}$ conditionally on $\boldsymbol{\Psi}$ by

$$\boldsymbol{\upsilon} = (\Sigma_{\hat{\mathbf{Z}}(\mathbf{x}^*),\hat{\mathbf{Z}}}, \Sigma_{\hat{\mathbf{Z}}(\mathbf{x}^*),\tilde{\mathbf{Z}}_1}, \ldots, \Sigma_{\hat{\mathbf{Z}}(\mathbf{x}^*),\tilde{\mathbf{Z}}_p}). \tag{5.31}$$

Hence, the distribution of $\hat{\mathbf{Z}}(\mathbf{x}^*)$ conditionally on the observed data $\hat{\mathbf{Z}}, \tilde{\mathbf{Z}}_{1:p}$ and $\boldsymbol{\Psi}$ is normal with mean $\hat{\boldsymbol{\mu}}(\mathbf{x}^*) + \boldsymbol{\upsilon}^\top \check{\boldsymbol{\Sigma}}^{-1}(\check{\mathbf{Z}} - \check{\boldsymbol{\mu}})$ and covariance $\boldsymbol{\Sigma}_{\hat{\mathbf{Z}}(\mathbf{x}^*)} - \boldsymbol{\upsilon}^\top \check{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\upsilon}$.

The posterior predictive distribution of $\hat{\mathbf{Z}}(\mathbf{x}^*)$ can be obtained as follows

$$p(\hat{\mathbf{Z}}(\mathbf{x}^*)|\hat{\mathbf{Z}}, \tilde{\mathbf{Z}}_{1:p}) = \int p(\hat{\mathbf{Z}}(\mathbf{x}^*)|\hat{\mathbf{Z}}, \tilde{\mathbf{Z}}_{1:p}, \boldsymbol{\Psi}) p(\boldsymbol{\Psi}|\hat{\mathbf{Z}}, \tilde{\mathbf{Z}}_{1:p}) d\boldsymbol{\Psi}. \tag{5.32}$$

After having obtained N simulations from the posterior distribution of $\boldsymbol{\Psi}$, it is then possible to approximate the integral in (5.32) by the following Rao-Blackwellized estimator:

$$p(\hat{\mathbf{Z}}(\mathbf{x}^*)|\hat{\mathbf{Z}}, \tilde{\mathbf{Z}}_{1:p}) \approx \frac{1}{N} \sum_{l=1}^{N} p(\hat{\mathbf{Z}}(\mathbf{x}^*)|\hat{\mathbf{Z}}, \tilde{\mathbf{Z}}_{1:p}, \boldsymbol{\Psi}^{(l)}), \tag{5.33}$$

where $\boldsymbol{\Psi}^{(l)}$ denotes the l-th simulated value from the posterior distribution of $\boldsymbol{\Psi}$.

5.4 Ensemble Modelling of Temperatures in the Pacific Northwest

5.4.1 Data Description

Recall the Probcast data introduced in Section 4.3.1. They include 48-hour forecasts of surface level temperature data initialized at midnight Coordinated Universal Time (UTC). The data come from the UW mesoscale short-range ensemble system for the Pacific northwestern area and corresponds to a five-member short-range ensemble consisting of different runs of the MM5 model, namely AVN, GEM, ETA, NGM, and NOGAPS. In these data, the grid-scaled deterministic model outputs had previously been interpolated to the locations of the monitoring stations by the Probcast group. The Figure 5.1 illustrates the locations of the 120 stations used in this chapter for illustration of the BEM methodology.

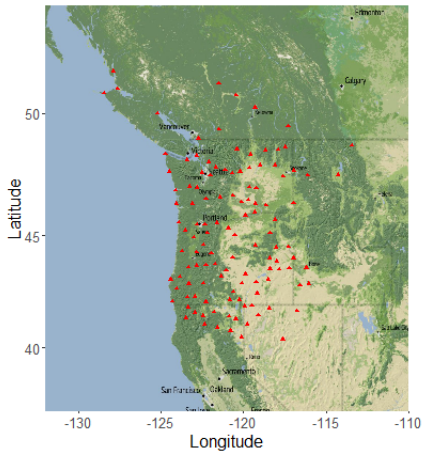


Figure 5.1: Map of the 120 stations used for illustration of the BEM methodology. Map created using `R` library `ggmap` with tiles by Stamen Design, under CC BY 3.0, and data by OpenStreetMap, under CC BY SA.

5.4.2 Inference

In Appendix B, for validation of our approximate inference computation, we discuss the application of the INLA-SPDE methodology for the BEM model in an artificial data setting. In this section, we discuss our findings for the Probcast Group data set.

Model Description

Since the deterministic model outputs had already been interpolated to the locations of the monitoring stations by the Probcast group, the BEM model can thus be simplified to

$$\hat{\mathbf{Z}}|\mathbf{Z},\boldsymbol{\Psi} \sim \mathcal{N}(\mathbf{Z},\sigma_e^2\mathbf{I}_n) \tag{5.34}$$

$$\tilde{\mathbf{Z}}_j|\mathbf{Z},\boldsymbol{\Psi} \sim \mathcal{N}(a_j\mathbf{1}_n + b_j\mathbf{Z},\sigma_{\delta_j}^2\mathbf{I}_n) \tag{5.35}$$

$$\mathbf{Z}|\boldsymbol{\Psi} \sim \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma}) \tag{5.36}$$

$$\boldsymbol{\Psi} \sim p(\boldsymbol{\Psi}), \tag{5.37}$$

for $j = 1, \ldots, p$, and $p(\mathbf{\Psi})$ denotes the prior distribution for $\mathbf{\Psi}$. In the above, $\mathcal{N}$ denotes a normal distribution with given mean vector and covariance matrix.

Figure 5.2 illustrates the overall correlations between measurements and model outputs. Note that the correlation is stronger within members of the ensemble. Due to the high correlations, a model that links the measurements with the model outputs via a common underlying field, such as the BEM model, seems like a sensible choice.

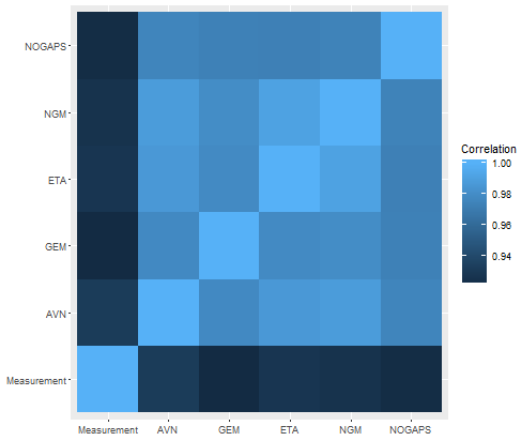


Figure 5.2: Pearson's correlation coefficients between measurements and model outputs.

Mean Description

In a similar manner as described in Section 4.5.1 for a proxy data set covering the same time frame and similar spatial domain, we assume a spatial mean that depends on the interaction between latitude (s_1) and longitude (s_2), and includes elevation ($h(\mathbf{s})$). The mean function can thus be summarized as

$$\mu(\mathbf{s}) = \beta_0 + \beta_1 s_1 + \beta_2 s_2 + \beta_3 s_1 s_2 + \beta_4 h(\mathbf{s}). \tag{5.38}$$

Prior Specifications

For model completeness, in order to carry out the inference procedure for the BEM for the Probcast data set, we describe our independent and vague prior specifications below. For numerical stability, we specify priors for precision parameters (inverse of the variance) in a logarithmic scale.

All notation used below for normal priors refer to mean and precision, whereas we refer to a Log-gamma as simply the logarithm of a Gamma distribution.

- For the mean parameters α, β_1, β_2, β_3, β_4, and calibration parameters a_j and b_j, for $j = 1, \ldots, 5$, we specified a $\mathcal{N}(0, 0.01)$ prior. This yields fairly noninformative priors for these parameters.
- For $\log(\sigma_{\delta_j}^{-2})$, $j = 1, \ldots, 5$, we specified a Log-gamma$(0.01, 0.01)$ prior.
- For $\log(\sigma_e^{-2})$, we specified a Log-gamma$(1, 0.01)$ prior.
- For $\log(\sigma)$, we specify a $\mathcal{N}(0, 0.1)$ prior, for $\log(\kappa)$ a $\mathcal{N}(0, 1)$. We heuristically specify the prior for the spatial range as a fifth of the approximate domain diameter. This leads to a fairly vague prior specification for $\log(\sigma)$. As described in Lindgren and Rue (2015), for this heuristic choice, the precision 1 for the prior of $\log(\kappa)$ gives an approximate 95% prior probability for the range being shorter than the domain size.

BEM Implementation in `R-INLA`

In this section, we briefly discuss the computational implementation of the BEM model using the `INLA-SPDE` framework. Our approach is similar to Ruiz-Cárdenas et al. (2012), where we construct an artificial observational equation with "fake zero" data for the BEM model implementation. This technique has become common while implementing advanced models in INLA, as described in the Chapter 8 of Blangiardo and Cameletti (2015). More recently, Rue et al. (2016) provide a review of Bayesian computing us-

ing the `R-INLA` package, and Lindgren and Rue (2015) focus on the aspects of the SPDE implementation.

Firstly, we construct the following linear predictors

$$\xi_0(\mathbf{s}) = \mu(\mathbf{s}) + P(\mathbf{s}, \mathbf{s}_0)g(\mathbf{s}_0) \tag{5.39}$$

$$\xi_j(\mathbf{s}) = a_j + b_j\xi_0(\mathbf{s}), \tag{5.40}$$

for $j = 1, \ldots, m$, where $g(\mathbf{s}_0)$ represent the SPDE triangulated mesh based on a Matérn model, and $P(\mathbf{s}, \mathbf{s}_0)$ is a projector matrix. The projector matrix is responsible for projecting the process from the mesh vertices to the observed locations. Figure 5.3 illustrates the triangulation of the spatial domain for Feb 20th, as described in Section 5.3.1. The triangulation was performed similarly for the other days.

Using the `R-INLA` terminology (Rue et al., 2016), the ξ_0 linear predictor, is created so that it can be "copied" into the linear predictors ξ_j, associated with the deterministic model outputs. The "copy" feature of `R-INLA` is key in the BEM implementation, as it will allow us to link the underlying process to both measurements and model outputs observational models.

The $\xi_j(\mathbf{s})$ predictors are constructed in an independent and identically distributed model with low arbitrary precision. This means that, in practice, $\xi_j(\mathbf{s})$ are free to vary though essentially they will be restricted to the linear functional described above. This is possible by assuming a Gaussian likelihood with "fake zero" observations with a high fixed precision. Examples of this approach can be found in Ruiz-Cárdenas et al. (2012). More specifically, we can describe the BEM observational model as

$$\mathbf{0}(\mathbf{s}) = \xi_0(\mathbf{s}) + \boldsymbol{\epsilon}_0 - \hat{\mathbf{z}}(\mathbf{s}) \tag{5.41}$$

$$\tilde{\mathbf{z}}_j(\mathbf{s}) = \xi_j(\mathbf{s}) + \boldsymbol{\epsilon}_j, \tag{5.42}$$

for $j = 1, \ldots, m$, where $\mathbf{0}(\mathbf{s})$ denotes the "fake zero" observations, $\boldsymbol{\epsilon}_0$ represents an independent Normal random effect with zero mean and variance σ_e^2 and, similarly, $\boldsymbol{\epsilon}_j$ independent Normal random effects, each with zero mean and variance $\sigma_{\delta_j}^2$.

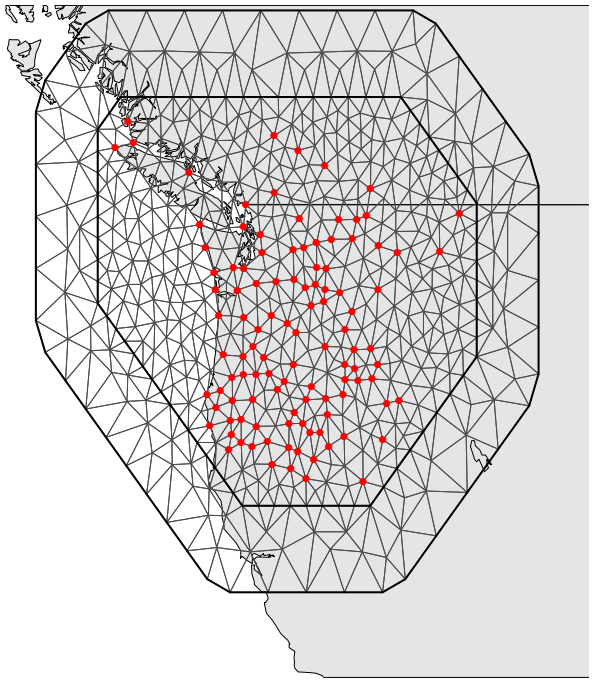


Figure 5.3: Triangulation for the BEM data available on Feb 20th. The mesh comprises of 591 edges and was constructed using triangles that have a minimum angle of 25, maximum edge length of 1° within the spatial domain and 2° in the extension domain. The maximum edges were chosen to be less than the approximate range of the process. The spatial domain was extended to avoid a boundary effect. The monitoring stations are highlighted in red.

Assessment of Calibration of the Model Outputs

An interesting feature of the BEM model is that by linking the deterministic model outputs process with the measurement process, it is then possible to quantify the uncertainty about these outputs. Of particular interest is therefore the calibration parameters (a_j and b_j) and variances $\sigma^2_{\delta_j}$ associated to each member of the ensemble, here $j = 1, \ldots, 5$.

Figure 5.5 illustrates the approximate marginal posterior distributions for these parameters for three selected days. The calibration parameters (a_j and b_j) are particularly useful to assess deviations from the assumed latent process. Figure 5.4 illustrates their variation over time. For the additive

calibration parameters (a_j), note the slight decreasing trend during colder periods followed by an increasing trend during warmer periods. On the other hand, the multiplicative calibration parameters (b_j) show a decreasing trend during warmer periods preceded by a slight increasing trend in colder periods.

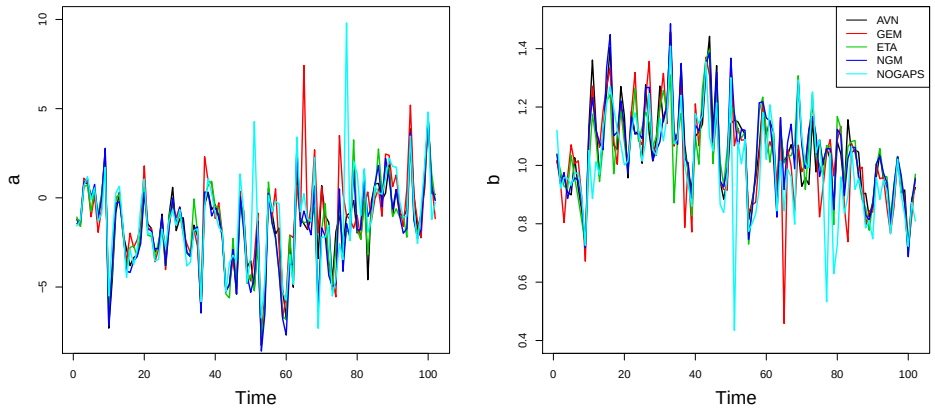


Figure 5.4: Posterior mean for the calibration parameters (a_j and b_j) for each member of the ensemble $j = 1, \ldots, 5$ across time (in days).

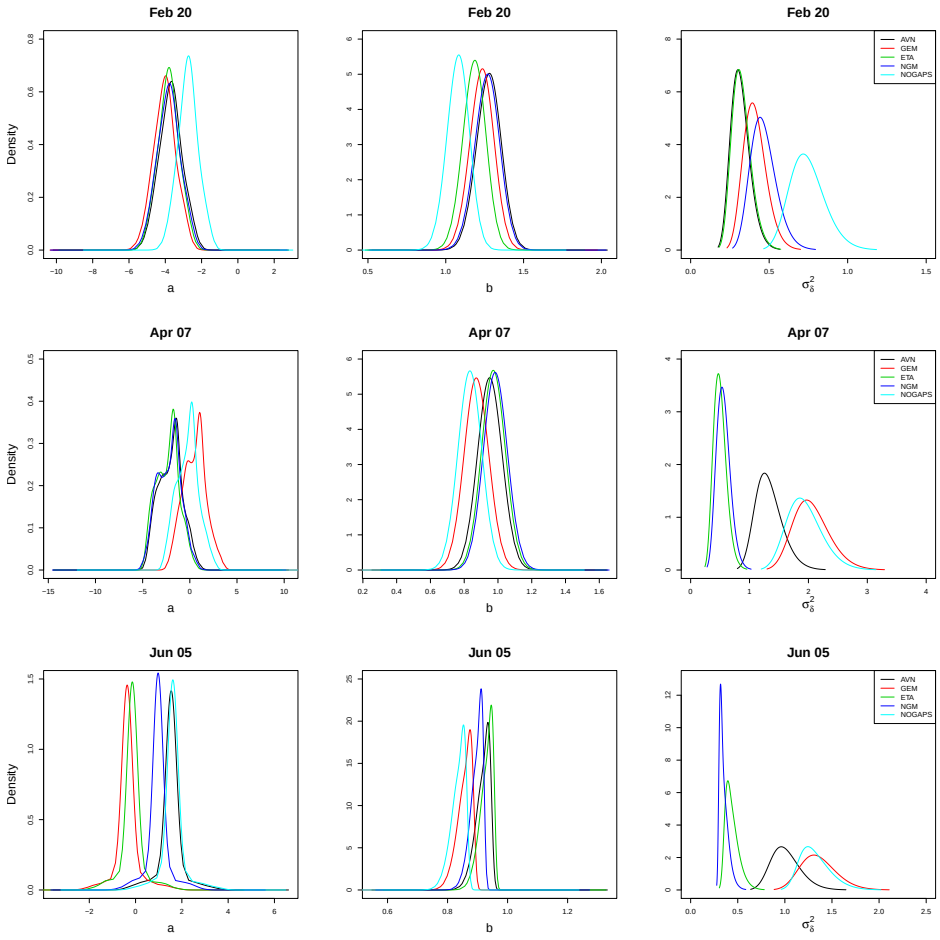


Figure 5.5: Approximate marginal posterior distributions for calibration parameters (a_j and b_j) and variances $\sigma^2_{\delta_j}$ for each member of the ensemble $j = 1, \ldots, 5$ for three selected days: February 20th, April 7th, and June 5th.

5.5 Discussion and Future Work

In this chapter, we introduced a scalable inference methodology alternative for the BEM using integrated nested Laplace approximations following Rue et al. (2009); Lindgren et al. (2011). In Chapter 6, we introduce a dynamic strategy that builds on the BEM model ability to accommodate time, with the objective of performing forecasting. Essentially we will apply

the BEM model sequentially over a set of training days. These posteriors will then be part of a mixture that will eventually be used for obtaining the predictive distributions for forecasting.

A limitation of the BEM model described in this Chapter is due to the fact that we assumed a Matérn covariance structure for the "true" underlying random field. From Chapter 4, we noted that it is crucial to handle non-stationarity when modelling temperatures in the Pacific Northwest. For future work, we would like to accommodate this into the BEM inference strategy.

As in Lindgren et al. (2011), this entails representing a Gaussian random field η as a solution to a SPDE with covariance parameters varying over space, and being written as

$$(\kappa^2(\mathbf{s}) - \Delta)^{\frac{\alpha}{2}}\{\tau(\mathbf{s})\eta(\mathbf{s})\} = \mathcal{W}(\mathbf{s}), \tag{5.43}$$

where τ models the variance of the process and is allowed to vary on space.

Chapter 6

Ensemble Forecaster

6.1 Contributions

In this chapter, we introduce a dynamic Bayesian ensemble model forecaster (DBEM), which essentially builds on the ability of the BEM model described Chapter 5 to accommodate time, with the objective of performing forecasting.

This general idea was originally introduced in Liu (2007), but the methodology was not fully developed due computational challenges, and hence no empirical assessment of the method was made. The main contribution of this chapter is in its critical assessment of the DBEM, as well as providing a detailed description of its strategy.

Following up on the inference for the BEM model based on an INLA framework described in the previous chapter, we offer a way to solve the computational challenges faced by Liu (2007) by making use of some approximations from the INLA framework. Moreover, we provide a comparison of its forecasting strengths with a Bayesian Model Averaging (BMA) alternative. The BMA is described in the following Section 6.2.

6.2 Bayesian Model Averaging

The main idea behind performing Bayesian model averaging (BMA) comes from realizing the need to take into account the uncertainty over the model choice in a particular setting. This idea flows quite naturally into a Bayesian framework. Instead of using a single model, the BMA relies on a mixing over multiple models, where the weights used for combining the quantities of interest are based on *posterior* model probabilities.

Consider ψ a quantity of interest, say, a predictive quantity, and let $M_1, \ldots, M_p$ denote the set of the possible models considered. The posterior probability for ψ is given by an average of the posterior distributions under each of the models considered is defined as follows

$$p(\psi|\mathcal{D}) = \sum_{j=1}^{p} p(\psi, |\mathcal{D}, M_j) p(M_j|\mathcal{D})$$

where $\mathcal{D}$ denotes the data available.

Empirically, this approach has proved superior to simply choosing a single "best" model based on some criteria. For instance, using a logarithmic scoring rule, Raftery et al. (1997) noted that the BMA had optimum predictive performance. A comprehensive review of the BMA can be found in Hoeting et al. (1999).

Consider now the situation where an ensemble of computer models outputs are available and there is interest in embedding them in a statistical model with the goal of obtaining probabilistic forecasts. The Bayesian model averaging forecast idea was introduced in Raftery et al. (2005), and it is based on a weighted average of individual deterministic outputs that constitute the ensemble, where the weights represent the individual forecast performance in a training period. The main idea is that there is a "best" model output, and they attempt to quantify the uncertainty about which member may be considered "best" using BMA.

Denoting by y as a weather quantity of interest and f_k the forecast associated with the k-th deterministic model in the ensemble, $k = 1, \ldots, p$, Raftery et al. (2005) apprise that the conditional distribution $g_k(y|f_k)$ can be interpreted as a conditional distribution of the data given the k-th deterministic model that has the best forecast performance in the ensemble. The BMA predictive model can therefore be written as

$$p(y|f_1, \ldots, f_k) = \sum_{k=1}^{p} w_k g_k(y|f_k), \tag{6.1}$$

where $\sum_{k=1}^{p} w_k = 1$ and $w_k = p(f_k|\mathcal{D})$ is the posterior probability of forecast

k being the best based on their predictive performance in a training period.

In particular, for the context of surface temperature forecasting, Raftery et al. (2005) use a normal distribution to approximate the conditional distribution $g_k(y|f_k)$, centred at linearly calibrated forecasts

$$Y|f_k \sim \mathcal{N}(a_k + b_k f_k, \sigma_0^2). \tag{6.2}$$

In contrast to the BEM model described in Section 5.2, they assume that the forecasts have been previously interpolated to the observation sites.

The BMA predictive mean is given by

$$\mathbb{E}[Y|f_1, \ldots, f_p] = \sum_{k=1}^{p} w_k(a_k + b_k f_k). \tag{6.3}$$

Note that the BMA estimation is done for one location at a time. For simplicity, we suppressed the spatial and temporal indexes in the equations above. The procedure is as follows. Raftery et al. (2005) first estimate separately the calibration parameters a_k and b_k for each member of the ensemble via least squares regression using the training data. Then, they proceed with the estimation of BMA weights w_k and the BMA variability σ_0^2 simultaneously for the ensemble via the Expectation Maximization (EM) algorithm for the training data. They optimize the estimation of σ_0 so that it optimizes the continuous ranked probability score (CRPS) for the training data by performing a numerical search over possible values of σ_0 centred at the maximum likelihood estimates, and keeping the other parameters fixed. The CRPS measures the difference between the predicted and true cumulative distributions, as follows

$$CRPS(F, y^*) = \int_{-\infty}^{\infty} [F(y^*) - \mathbf{1}_{y \geq y^*}]^2 d_y, \tag{6.4}$$

where F is the cumulative distribution function of the forecast distribution, y^* is the true value and $\mathbf{1}_{y \geq y^*}$ is a step function that attains the value 1 when $y \geq y^*$ and zero otherwise. The BMA is implemented in `R` in the `ensembleBMA` package (Fraley et al., 2013).

On the other hand, Berrocal et al. (2007) extend the BMA by taking into account the possibility of spatial correlation, unlike the basic BMA. Their spatial BMA strategy describes the predictive distribution for the whole field as a weighted average of multivariate normal distributions centred at linearly calibrated members of the ensemble. Similarly, Kleiber et al. (2011) provide a spatially adaptive extension called geostatistical model averaging by first estimating the parameters of the BMA model at each location and then interpolating the estimates using a geostatistical model.

In the following Section 6.3, we introduce our alternative strategy, based on applying the BEM of Chapter 5 for the purposes of forecasting.

6.3 Dynamic Bayesian Ensemble Forecaster

6.3.1 Decision Making Ideas

The motivation for introducing the dynamic Bayesian ensemble model forecaster (DBEM) comes from accommodating time into the BEM model described in Section 5.2 by borrowing ideas from a decision making model proposed by Bayarri and DeGroot (1989) to combine opinions from different experts.

Each of the k experts was assumed to have their own uncertainty about a certain parameter W quantified by the individual prior distributions denoted as $\pi_i(w)$ assigned by each expert i. In this decision making model, it was assumed that an executive would form their opinion about W through a weighted average of the experts' opinions. The executive's prior is

$$\pi(w) = \sum_{i=1}^{k} \alpha_i^{(0)} \pi_i(w), \text{ where } \sum_{i=1}^{k} \alpha_i^{(0)} = 1, \ \alpha_i^{(0)} \geq 0. \tag{6.5}$$

Furthermore, the k experts and the executive were assumed to jointly observe the value of a random variable X whose conditional distribution when $W = w$ is denoted as $f(\cdot|w)$. The posterior for expert i is

$$\pi_i^*(w|x) = \frac{\pi_i(w) f(x|w)}{p_i(x)}, \tag{6.6}$$

where $p_i(x)$ is the marginal distribution for expert i, that is,

$$p_i(x) = \int f(x|w)\pi_i(w)dw. \tag{6.7}$$

The marginal distribution for the executive is given by

$$p(x) = \sum_{i=1}^{k} \int \alpha_i^{(0)} f(x|w)\pi_i(w)dw. \tag{6.8}$$

Hence, the posterior for the executive is

$$\begin{aligned} \pi^*(w|x) &= \frac{\pi(w)f(x|w)}{p(x)} \\ &= \sum_{i=1}^{k} \frac{\alpha_i^{(0)} p_i(x)}{p(x)} \frac{\pi_i(w)f(x|w)}{p_i(x)} \\ &= \sum_{i=1}^{k} \alpha_i^{(1)} \pi_i^*(w|x). \end{aligned} \tag{6.9}$$

Therefore, the posterior for the executive is again a weighted average of posteriors ($\pi_i^*(w|x)$, $i = 1, \ldots, k$) for the k experts. The updated weight for expert i after observing $X = x$ is $\alpha_i^{(1)} = \alpha_i p_i(x)/p(x)$, which depends on their marginal distribution of X. Note that the expert's weight will increase if his/her marginal distribution of x is large.

In the following subsection, we describe the dynamic Bayesian ensemble melding forecaster, which borrows ideas from this decision making strategy. Essentially, the measurements $\hat{\mathbf{Z}}$ and the members of the ensemble $\tilde{\mathbf{Z}}_1, \ldots, \tilde{\mathbf{Z}}_p$ across space at a given time point are viewed as the X in this decision-making process. The first k time point measurements and model outputs are used to fit the BEM model described in Section 5.2, and thus giving the posterior distribution of $\mathbf{\Psi}$ for each of these k time points.

6.4 DBEM Forecaster

The dynamic Bayesian ensemble model forecaster borrows ideas from the decision making strategy described in Section 6.3.1, with the purpose of accommodating time in the BEM model, which is fundamentally a spatial model.

In order to do so, a set of k time point measurements and model outputs are used as a training set. These data are used to fit the BEM model described in Section 5.2, and obtain the posterior distribution of $\mathbf{\Psi}$ for each of these k time points. Cross-validation techniques can be used for choosing the number of time points. By combining these posterior distributions, an "executive" posterior distribution is obtained with the purpose of representing the uncertainty about $\mathbf{\Psi}$ after having observed data across the training days. This "executive" posterior can thus be viewed as a prior for the subsequent day which could then be used to obtain the predictive distribution for this future day.

6.4.1 A DBEM Forecaster Algorithm

In this section, we provide details about the DBEM forecasting strategy, described in Algorithm 1, with the objective of obtaining forecasts.

Following up on Section 5.3.2, recall that we introduced the joint distribution of $\hat{\mathbf{Z}}, \tilde{\mathbf{Z}}_{1:p}$ conditionally on the hyperparameters $\mathbf{\Psi}$ as a multivariate normal. Hence, the conditional distribution $\hat{\mathbf{Z}}_{k+1}|\mathbf{\Psi}^{(l)}, \tilde{\mathbf{Z}}_{1:p,k+1}$, for each sample l of $\mathbf{\Psi}$ required for Algorithm 1 can be obtained by using the properties of multivariate normal.

The main advantage of the DBEM is that it sidesteps the need to build time series models to incorporate a temporal component and instead smooths the sequence of posteriors by averaging them over time. It also builds upon the Bayesian ensemble model's ability to forecast temperatures at future times. In contrast to the BMA approach described in Section 6.2, another very desirable feature of the DBEM is that it also allows the calibration coefficients (a_k, b_k, for $k = 1, \ldots, p$) to change over time.

Algorithm 1 DBEM Forecaster

Let $i = 1, \ldots, k$ index a set of training days.

1. The first step is to initialize the weights. For instance, by setting $\alpha_i^{(0)} = \frac{1}{k}$, for each $i = 1, \ldots, k$, no preference is given to any time point in the training set.

2. Obtain M samples from $\pi_i^{(0)}(\mathbf{\Psi}|\hat{\mathbf{Z}}_i, \tilde{\mathbf{Z}}_{1:p,i})$, i.e., the posterior for each training day i, where $\tilde{\mathbf{Z}}_{1:p,i} = (\tilde{\mathbf{Z}}_{1,i}, \ldots, \tilde{\mathbf{Z}}_{p,i})$, and $\tilde{\mathbf{Z}}_{p,i}$ denotes the data available for the p-th member of the ensemble at the i-th training day.

3. Update weights and get samples from the "executive" posterior described as follows

$$\pi^{(1)}(\mathbf{\Psi}|\hat{\mathbf{Z}}_k, \tilde{\mathbf{Z}}_{1:p,k}) = \sum_{i=1}^{k} \alpha_i^{(1)} \pi_i^{(0)}(\mathbf{\Psi}|\hat{\mathbf{Z}}_i, \tilde{\mathbf{Z}}_{1:p,i}), \tag{6.10}$$

   where

$$\alpha_i^{(1)} = \frac{\alpha_i^{(0)} p_i(\hat{\mathbf{Z}}_i, \tilde{\mathbf{Z}}_{1:p,i})}{\sum_{i=1}^{k} \alpha_i^{(0)} p_i(\hat{\mathbf{Z}}_i, \tilde{\mathbf{Z}}_{1:p,i})}. \tag{6.11}$$

4. The "executive" posterior is used as a prior in order to obtain the following predictive distribution:

$$\begin{aligned} & f(\hat{\mathbf{Z}}_{k+1}|\tilde{\mathbf{Z}}_{1:p,k+1}) && (6.12)\\ = & \int f(\hat{\mathbf{Z}}_{k+1}|\mathbf{\Psi}, \tilde{\mathbf{Z}}_{1:p,k+1}) \pi^{(2)}(\mathbf{\Psi}) d\mathbf{\Psi} && (6.13)\\ \approx & \frac{1}{L} \sum_{l=1}^{L} f(\hat{\mathbf{Z}}_{k+1}|\mathbf{\Psi}^{(l)}, \tilde{\mathbf{Z}}_{1:p,k+1}), && (6.14) \end{aligned}$$

   where $\mathbf{\Psi}_1^{(l)}$, $l = 1, \ldots, M$ are samples from the executive posterior.

In addition, the DBEM is also able to accommodate spatial correlation since the posterior distributions for the training days are obtained via the Bayesian ensemble model, which is essentially a spatial model. In contrast, the BMA model strategy is done for one location at a time, as described in Section 6.2.

The posteriors in item 2 of the DBEM algorithm can be approximated efficiently using the `INLA-SPDE` approach described in Section 5.3.1, thus avoiding the computational burden associated with MCMC strategies. Moreover, significant computational benefits come from efficiently obtaining the marginal likelihoods $p_i(\hat{\mathbf{Z}}_i, \tilde{\mathbf{Z}}_{1:p,i})$ from `R-INLA` as described in (3.2).

A limitation of this methodology is due to the fact that the marginal log-likelihood approximations may be difficult to compare in situations where the number of stations vary significantly among the training days.

The marginal log-likelihoods will usually be very negative, which might yield some difficulty in differentiating the weights that measure the contribution of each training day in the mixture. This poor mixing will tend to happen so long as at least one log-likelihood is significantly greater than the rest, and will ultimately dominate the mixing weights. Moreover, another limitation of the methodology described in Liu (2007) is that it does not take into consideration the order of the observations. Instead, it heavily relies on the quality of the data in each of the training days.

In order to overcome the poor mixing when obtaining the samples from the "executive" posterior, we scale the marginal log-likelihoods by the average of the marginal log-likelihoods in the training set. The weights in part 3 of Algorithm 1 are now written as

$$\alpha_i^{(1)} = \frac{\alpha_i^{(0)} \exp\{m^{-1} \log p_i(\hat{\mathbf{Z}}_i, \tilde{\mathbf{Z}}_{1:p,i})\}}{\sum_{i=1}^{k} \alpha_i^{(0)} \exp\{m^{-1} \log p_i(\hat{\mathbf{Z}}_i, \tilde{\mathbf{Z}}_{1:p,i})\}}, \tag{6.15}$$

where $m = |\sum_{i=1}^{k} \log p_i(\hat{\mathbf{Z}}_i, \tilde{\mathbf{Z}}_{1:p,i})/k|$ is the absolute value of the average of all the marginal log-likelihoods in the training set. This strategy maintains the order of influence of the marginal log-likelihood among the training days, but deflates/inflates the weights to insure a better mixing.

In the following section, we describe an empirical assessment of the DBEM methodology for obtaining forecasts. We compare it with the BMA methodology described in Section 6.2.

6.5 An Empirical Assessment of the DBEM

In this section, we again refer to Probcast data introduced in Section 4.3.1. To illustrate the DBEM methodology, Figure 5.1 contains the map of the 120 stations considered in this study. Our goal is to perform an empirical assessment of the DBEM methodology by comparing it with the BMA approach described in Section 6.2. Recall that the number of available stations varies across the different time points, as noted in Figure 4.2.

In Section 6.5.1, we discuss our findings based on sequentially fitting a BEM (as illustrated in Section 5.4) among the training days, and using a DBEM strategy to obtain forecasts one day ahead. Moreover, in Section 6.5.2 we use our knowledge of the behaviour of the temperature fields in the Pacific Northwest described in Section 4.5.1 in an attempt to alleviate the influence of localized weather.

6.5.1 Forecasting Temperature

The first required step for the application of both the DBEM and the BMA entails specifying the number of training days. It would be advantageous to use a shorter training period in order to adapt rapidly to the changes in weather patterns. For the same data set used in this thesis, Raftery et al. (2005) noted that the overall root mean squared forecast error (RMSFE) and the mean absolute error (MAE) of the BMA decreased substantially as the number of training days increases, up to 25 days, with little change in performance beyond that. Because of this, we began with a training set of 25 days for both methods. We later perform cross-validation and investigate the effect of varying the number of training days in the forecasts.

Table 6.2 contains the forecasting summaries for three selected days: Feb 20th, Apr 7th and June 5th. Note that while observing a smaller RMSFE and MAE, and higher empirical coverages of the 95% credible intervals (CIs), the average length of the CIs for the BMA are generally larger than those for the DBEM. We believe this is due to the strength that the DBEM borrows from neighbouring sites at which a forecast is being made. We intend to explore this issue in future work. Possibly due to this, note in Figure 6.2

that DBEM outperformed the BMA for some of the stations.

Table 6.1: Forecasting summaries for three selected days February 20th, April 7th and June 5th, using a training set of 25 days. Summaries include the root mean squared forecast error (RMSFE), mean absolute error (MAE), the empirical coverage and the average length of the 95% credible intervals (CI) for the different methods considered: the dynamic Bayesian ensemble model and the Bayesian model averaging (BMA). There are a total of 109 available stations on Feb 20th, and a total of 105 on Apr 7th and June 5th.

Method	Selected Days	RMSFE (°C)	MAE (°C)	Empirical coverage	Average length of 95% CI (°C)
DBEM	Feb 20th	2.757	2.291	0.954	10.826
	Apr 7th	3.260	2.752	0.895	10.657
	Jun 5th	4.524	3.675	0.848	12.331
BMA	Feb 20th	2.793	2.224	0.982	12.496
	Apr 7th	3.304	2.708	0.943	12.121
	Jun 5th	2.452	1.973	1.000	13.217

The overall forecasting summaries across all time points can be found in Table 6.2. The average length of the CIs for the BMA are generally larger than those for the DBEM, but the BMA outperformed the DBEM as measured by its smaller RMSFE and MAE, and higher empirical coverages of the 95% credible intervals.

In order to get a better understanding about the forecasting ability of the different methods across time, in Figure 6.1 we illustrate the RMSFEs split by month. We also describe the performance statistics in Table 6.3. Note that the DBEM outperformed the BMA during the months of February and March, but significantly underperformed it for the month of June.

For cross-validation purposes, we also varied k to guide us in the choice of the number of training days. In Figure 6.3 we illustrate the RMSFEs obtained for the month of June in particular. Note that there is a minor improvement in performance for the BMA. For the DBEM, however, the number of training days does not seem to be a significant factor for performance improvement.

In fact, from Algorithm 1, one may note that the forecasting performance will be highly influenced by the quality of the posterior samples of the "executive" posterior. Thus, a limitation of the DBEM is that it will tend to underperfom under high uncertainty *a posteriori*. In particular, from the the 95% CIs in Figure 6.5, we note increases in uncertainty for the mean parameters later in the study. This may explain the observed decrease in performance of the DBEM for the month of June.

Recall that in Chapter 5 we mentioned a limitation of the BEM model due to its assumed Matérn covariance structure. This may degrade the DBEM's performance over some time periods, since it limits the ways in which strengths can be borrowed over space. The decrease in performance is particularly exacerbated in June, when there is more discrepancy in temperatures, as seen in Figure 6.4 and Table 6.4.

Table 6.2: Forecasting summaries across all available time points using a training set of 25 days. There are a total of 77 time points. Summaries include the root mean squared forecast error (RMSFE), mean absolute error (MAE), the empirical coverage and the average length of the 95% credible intervals for the different methods considered: the dynamic Bayesian ensemble model (DBEM) and the Bayesian model averaging (BMA).

Method	RMSFE (°C)	MAE (°C)	Empirical coverage	Average length of 95% CI (°C)
DBEM	4.003	3.347	0.823	11.240
BMA	3.023	2.381	0.942	12.051

Table 6.3: Forecasting summaries across the different months, using a training set of 25 days. Summaries include the root mean squared forecast error (RMSFE), mean absolute error (MAE), the empirical coverage and the average length of the 95% credible intervals (CI) for the different methods considered: the dynamic Bayesian ensemble model and the Bayesian model averaging (BMA).

Method	Selected Days	RMSFE (°C)	MAE (°C)	Empirical coverage	Average length of 95% CI (°C)
DBEM	February	2.542	2.038	0.965	10.922
	March	3.149	2.542	0.885	10.241
	April	3.816	3.101	0.855	11.009
	May	3.247	2.588	0.918	11.745
	June	6.084	5.354	0.617	12.025
BMA	February	2.568	2.038	0.983	12.385
	March	3.245	2.587	0.913	11.339
	April	3.554	2.736	0.912	12.506
	May	2.916	2.279	0.960	13.151
	June	2.795	2.217	0.954	11.454

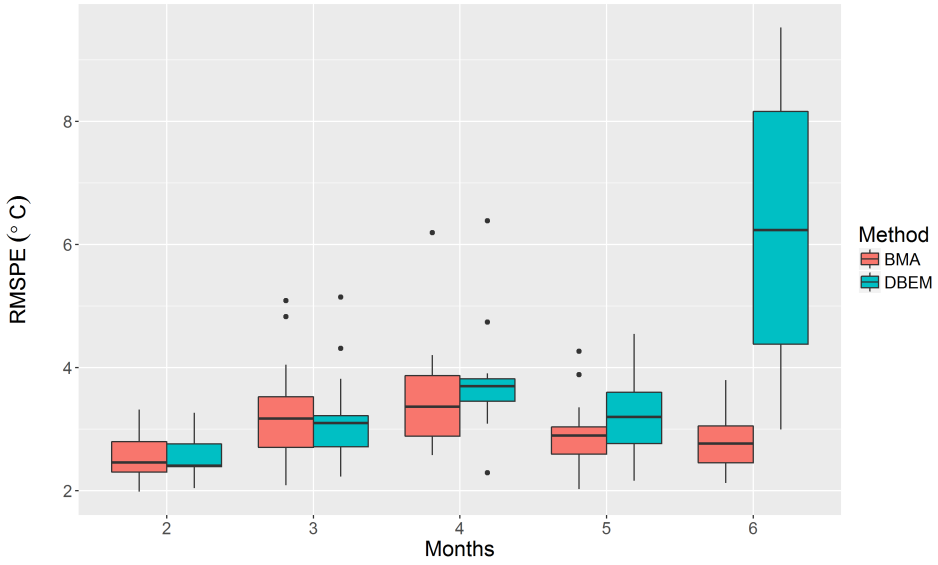


Figure 6.1: Mean squared forecast error (MSFE) across space for forecasts from February 20th to June 30th using the dynamic Bayesian ensemble model (DBEM) and the Bayesian model averaging (BMA). Both methods assumed a training set of 25 days.

(a) Feb 20th

(b) Apr 7th

(c) Jun 5th

Figure 6.2: 95% credible intervals of the forecasts for days Feb 20th, Apr 7th and June 5th for the different methods considered: the dynamic Bayesian ensemble model (DBEM) and the Bayesian model averaging (BMA). The number of stations where the forecasts were obtained were 109, 105 and 105, respectively. The dots represent the true measurement of temperature across the available stations for the selected days. To facilitate visualization, we also coloured the dots based on the different methodologies. Note the overall larger error bars observed for the BMA.

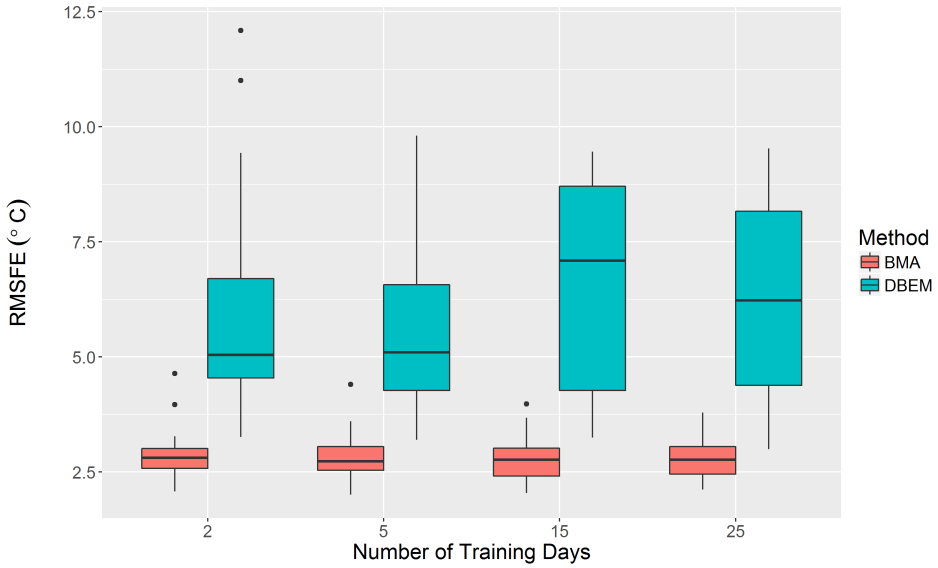


Figure 6.3: Mean squared forecast error (°C), MSFE, across space for forecasts across the month of June for different number of training days using the dynamic Bayesian ensemble model (DBEM) and the Bayesian model averaging (BMA).

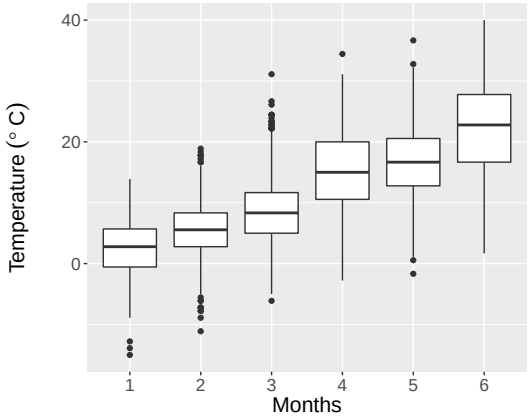


Figure 6.4: Monthly boxplots of observed temperatures over space.

Table 6.4: Summary statistics for the temperature measurements (°C) over space.

Month	Average (°C)	Std. Dev. (°C)
January	2.581	4.369
February	5.531	3.982
March	8.433	5.187
April	14.992	5.858
May	16.778	5.642
June	22.374	7.111

6.5.2 Forecasting Temperature Anomalies

In this subsection, we use our knowledge of the behaviour of the temperature fields in the Pacific Northwest described in Section 4.5.1 in an attempt to alleviate the influence of localized weather, particularly in warmer periods as seen in the previous Section 6.5.1. Here, we obtain the temperature anomalies by taking out the spatial mean fitted via least squares. We then proceed with forecasting these anomalies, and adding the fitted spatial mean back to obtain the forecasts in an observed temperature scale.

Table 6.5 contains the forecasting summaries for three selected days: Feb 20th, Apr 7th and June 5th. Comparing these results with those in Table 6.2, note the reduction in the RMSFE for the DBEM for June 5th. Figure 6.6 illustrates the 95% CIs of the forecasts for the selected days across the different stations.

An overall assessment across all time points can be found in Table 6.6 as well as split by month in Table 6.7. In Figure 6.7 we illustrate the RMSFEs split by month. Note that the DBEM still portrays a significantly poorer performance than the BMA for the month of June, though compared Section 6.5.1, we see an improvement in the RMSFE.

Since in Section 6.5.2 the mean parameters were estimated separately for all the training days, these parameters are thus varying with time and we are implicitly incorporating time there. This could help explain why this new approach based on anomalies did not yield a substantial improvement in the results.

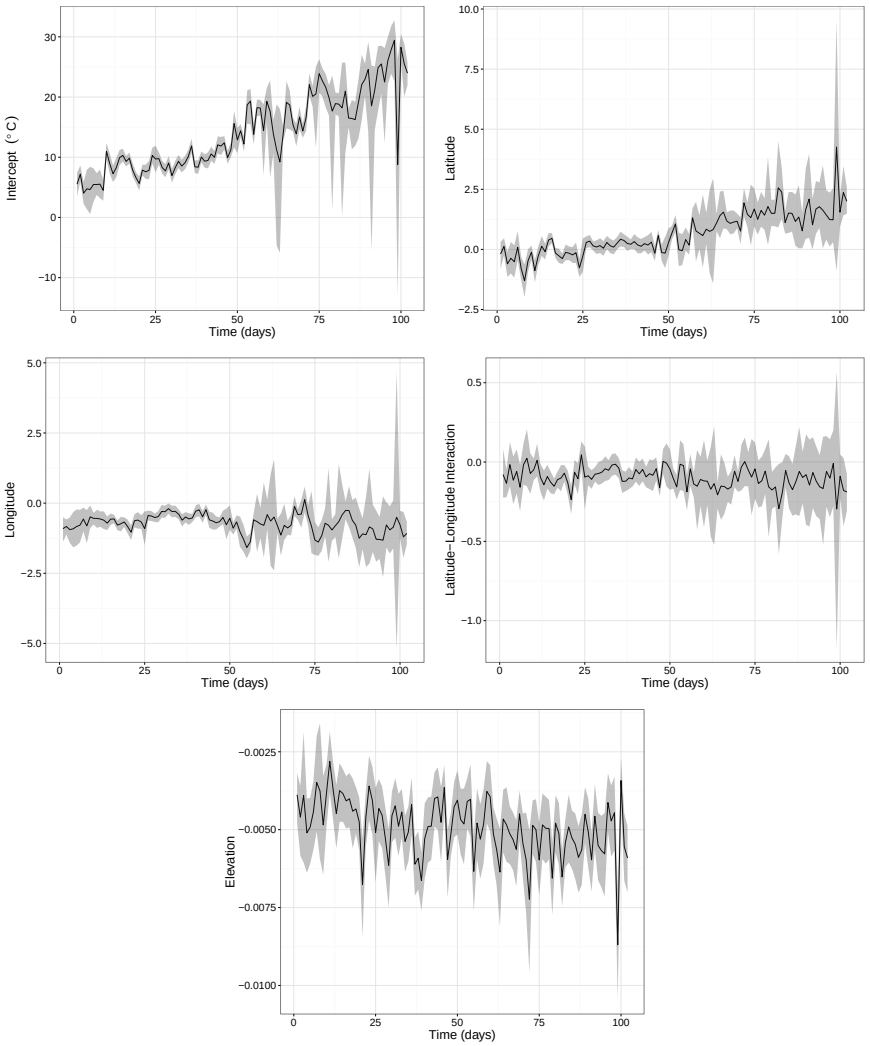
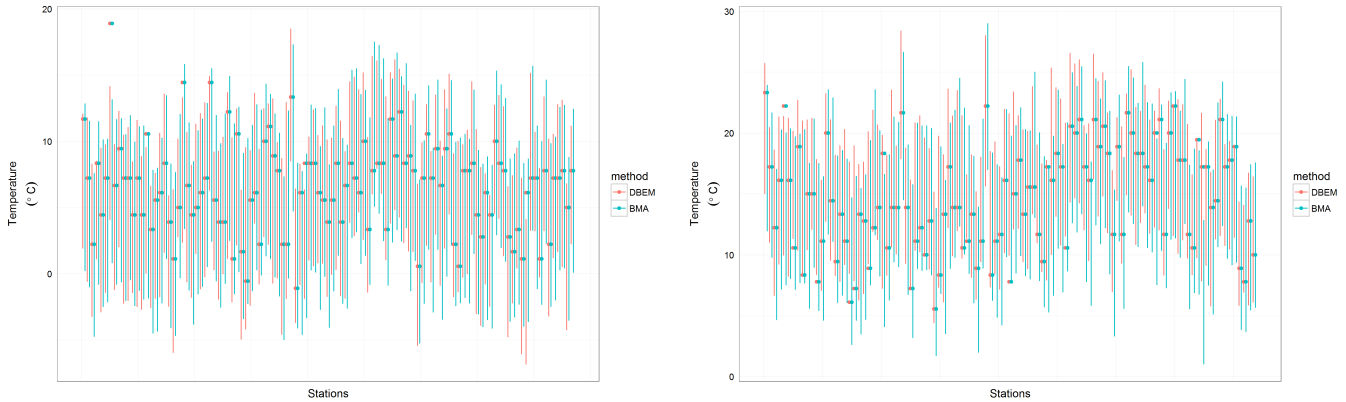


Figure 6.5: Posterior means (solid line) for the mean parameters of the underlying random field over time. The gray shaded region represent the 95% credible intervals. Note the increase in uncertainty for later days in the series.

Table 6.5: Forecasting summaries for three selected days Feb 20th, Apr 7th and Jun 5th, using a training set of 25 days. Summaries include the root mean squared forecast error (RMSFE), mean absolute error (MAE), the empirical coverage and the average length of the 95% credible intervals (CI) for the different methods considered: the dynamic Bayesian ensemble model and the Bayesian model averaging (BMA). There are a total of 109 available stations on Feb 20th, and a total of 105 on Apr 7th and June 5th.

Method	Selected Days	RMSFE (°C)	MAE (°C)	Empirical coverage	Average length of 95% CI (°C)
DBEM	Feb 20th	2.612	2.063	0.954	10.896
	Apr 7th	3.045	2.430	0.895	10.476
	Jun 5th	3.520	2.701	0.867	11.404
BMA	Feb 20th	2.793	2.224	0.982	12.496
	Apr 7th	3.304	2.708	0.943	12.121
	Jun 5th	2.452	1.973	1.000	13.217



(a) Feb 20th

(b) Apr 7th

(c) June 5th

Figure 6.6: 95% credible intervals of the forecasts for days Feb 20th, Apr 7th and June 5th for the different methods considered: the dynamic Bayesian ensemble model (DBEM) and the Bayesian model averaging (BMA). The number of stations where the forecasts were obtained were 109, 105 and 105, respectively. The dots represent the true measurement of temperature across the available stations for the selected days. To facilitate visualization, we also coloured the dots based on the different methodologies. Note the overall larger error bars observed for the BMA.

Table 6.6: Forecasting summaries across time using a training set of 25 days. Summaries include the root mean squared forecast error (RMSFE), mean absolute error (MAE), the empirical coverage and the average length of the 95% credible intervals for the different methods considered: the dynamic Bayesian ensemble model (DBEM) and the Bayesian model averaging (BMA).

Method	RMSFE (°C)	MAE (°C)	Empirical coverage	Average length of 95% CI (°C)
DBEM	3.647	2.995	0.842	10.812
BMA	3.023	2.381	0.942	12.051

Table 6.7: Forecasting summaries across the different months, using a training set of 25 days. Summaries include the root mean squared forecast error (RMSFE), mean absolute error (MAE), the empirical coverage and the average length of the 95% credible intervals (CI) for the different methods considered: the dynamic Bayesian ensemble model and the Bayesian model averaging (BMA).

Method	Selected Days	RMSFE (°C)	MAE (°C)	Empirical coverage	Average length of 95% CI (°C)
DBEM	February	2.643	2.103	0.948	10.909
	March	3.177	2.577	0.878	10.146
	April	3.551	2.822	0.882	10.937
	May	3.320	2.659	0.908	11.597
	June	4.805	4.110	0.691	10.702
BMA	February	2.568	2.038	0.983	12.385
	March	3.245	2.587	0.913	11.340
	April	3.553	2.736	0.913	12.506
	May	2.915	2.279	0.961	13.151
	June	2.795	2.217	0.954	11.454

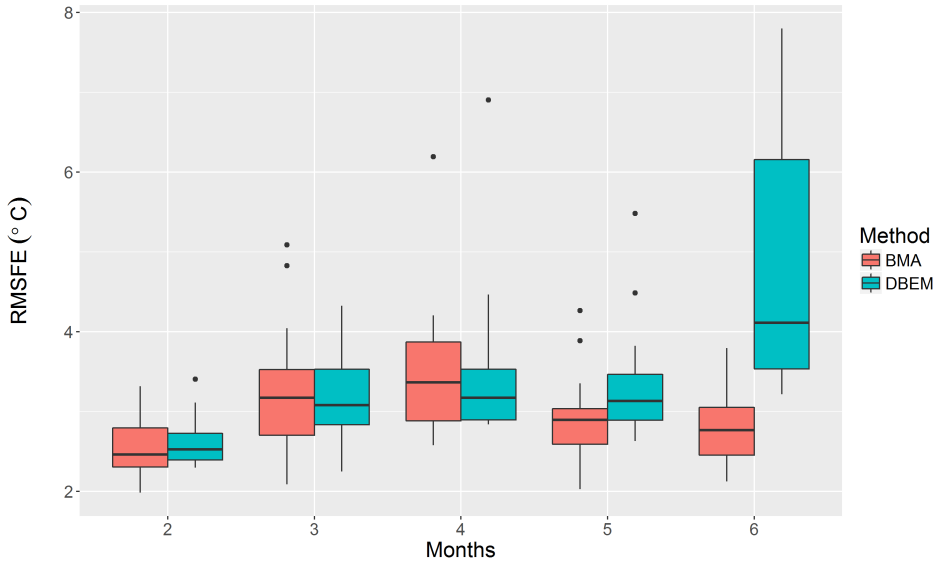


Figure 6.7: Mean squared forecast error (°C), MSFE, across space for forecasts from February 20th to June 30th using the dynamic Bayesian ensemble model (DBEM) and the Bayesian model averaging (BMA). Both methods assumed a training set of 25 days.

6.6 Discussion and Future Work

In this chapter, we introduced a dynamic alternative for the BEM which allows us to obtain forecasts, despite the fact that the BEM is a spatial model. The DBEM is a promising methodology as it outperformed the BMA for some time periods. A limitation of this methodology carries over from the previous chapter and is due to its assumed stationary covariance structure, which ultimately contributed to the decrease in performance for the DBEM particularly over warmer periods, where the discrepancy in temperature measurements is larger. Following up on this would require handling non-stationarity in the `INLA-SPDE` modelling.

Additionally, we discussed that the DBEM methodology requires the computation of mixing weights, which are based on normalized marginal likelihoods across a training set. Difficulties were encountered in differentiating amongst these weights when at least one log-likelihood is significantly higher than the rest.

To deal with this issue, we propose an alternative methodology that is the subject of current research. We describe the proposed methodology in Algorithm 2. The main difference is in the added step of first mixing the posteriors on the training set with equal weights, which is then viewed as a prior for the subsequent day. As a consequence, the mixing weights used to yield the "executive" posterior are based on individual marginal likelihoods of each of the training days but considering the data at a future time. Samples from this "executive" posterior are used to obtain an approximation of the predictive distribution at the subsequent time, and ultimately the forecasts.

Algorithm 2 Modified DBEM Forecaster

Let $i = 1, \ldots, k$ index a set of training days.

1. The first step is to initialize the weights, by setting $\alpha_i^{(0)} = \frac{1}{k}$, for each $i = 1, \ldots, k$. At this stage, no preference is given to any time point in the training set.

2. Obtain samples from $\pi_i(\mathbf{\Psi}|\hat{\mathbf{Z}}_i, \tilde{\mathbf{Z}}_{1:p,i})$, i.e., the posterior for each training day i, where $\tilde{\mathbf{Z}}_{1:p,i} = (\tilde{\mathbf{Z}}_{1,i}, \ldots, \tilde{\mathbf{Z}}_{p,i})$, and $\tilde{\mathbf{Z}}_{p,i}$ denotes the data available for the p-th member of the ensemble at the i-th training day. These will now represent "experts" posteriors.

3. A baseline "executive" prior for the subsequent day is based on a weighted average of the "experts" posteriors among the training days

$$\pi^{(0)}(\mathbf{\Psi}|\hat{\mathbf{Z}}_{1:k}, \tilde{\mathbf{Z}}_{1:p,1:k}) = \sum_{i=1}^{k} \alpha_i^{(0)} \pi_i(\mathbf{\Psi}|\hat{\mathbf{Z}}_i, \tilde{\mathbf{Z}}_{1:p,i}). \tag{6.16}$$

4. The next step will require samples from the "executive" posterior described below.

$$\begin{aligned} & \pi^{(1)}(\mathbf{\Psi}|\hat{\mathbf{Z}}_{k+1}, \tilde{\mathbf{Z}}_{1:p,k+1}) \\ = & \frac{f(\hat{\mathbf{Z}}_{k+1}, \tilde{\mathbf{Z}}_{1:p,k+1}|\mathbf{\Psi})\pi^{(0)}(\mathbf{\Psi}|\hat{\mathbf{Z}}_{1:k}, \tilde{\mathbf{Z}}_{1:p,1:k})}{p(\hat{\mathbf{Z}}_{k+1}, \tilde{\mathbf{Z}}_{1:p,k+1})} \\ = & \sum_{i=1}^{k} \frac{\alpha_i^{(0)} p_i(\hat{\mathbf{Z}}_{k+1}, \tilde{\mathbf{Z}}_{1:p,k+1})}{p(\hat{\mathbf{Z}}_{k+1}, \tilde{\mathbf{Z}}_{1:p,k+1})} \frac{f(\hat{\mathbf{Z}}_{k+1}, \tilde{\mathbf{Z}}_{1:p,k+1}|\mathbf{\Psi})\pi_i(\mathbf{\Psi}|\hat{\mathbf{Z}}_i, \tilde{\mathbf{Z}}_{1:p,i})}{p_i(\hat{\mathbf{Z}}_{k+1}, \tilde{\mathbf{Z}}_{1:p,k+1})} \\ = & \sum_{i=1}^{k} \alpha_i^{(1)} \pi_i(\mathbf{\Psi}|\hat{\mathbf{Z}}_{k+1}, \tilde{\mathbf{Z}}_{1:p,k+1}), \end{aligned} \tag{6.17}$$

   where the updated weights are given by

$$\alpha_i^{(1)} = \frac{\alpha_i^{(0)} p_i(\hat{\mathbf{Z}}_{k+1}, \tilde{\mathbf{Z}}_{1:p,k+1})}{\sum_{i=1}^{k} \alpha_i^{(0)} p_i(\hat{\mathbf{Z}}_{k+1}, \tilde{\mathbf{Z}}_{1:p,k+1})}. \tag{6.18}$$

5. Finally, the forecasts are given by:

$$\begin{aligned} & f(\hat{\mathbf{Z}}_{k+2}|\tilde{\mathbf{Z}}_{1:p,k+2}) \\ = & \int f(\mathbf{Z}_{k+2}|\mathbf{\Psi}, \hat{\mathbf{Z}}_{k+1}, \tilde{\mathbf{Z}}_{1:p,k+1})\pi^{(1)}(\mathbf{\Psi}|\hat{\mathbf{Z}}_{k+1}, \tilde{\mathbf{Z}}_{1:p,k+1})d\mathbf{\Psi} \quad (6.19) \\ \approx & \frac{1}{L}\sum_{l=1}^{L} f(\hat{\mathbf{Z}}_{k+2}|\mathbf{\Psi}^{(l)}, \hat{\mathbf{Z}}_{k+1}, \tilde{\mathbf{Z}}_{1:p,k+1}), \qquad 88 \end{aligned}$$

   where $\mathbf{\Psi}_1^{(l)}$, $l = 1, \ldots, M$ correspond to samples from $\pi^{(1)}(\mathbf{\Psi}|\hat{\mathbf{Z}}_{k+1}, \tilde{\mathbf{Z}}_{1:p,k+1})$.

Chapter 7

Determinantal Point Processes

In this chapter, we provide an overview of determinantal point processes, including definitions and sampling strategies. The main purpose of this chapter is to motivate the potential use of these processes in the design of monitoring networks, and ultimately provide a background for Chapter 8.

7.1 Motivation

Determinantal point processes (DPPs) are repulsion point processes and very appealing in practice due to their exact inference properties. DPPs have been explored in random matrix theory since the 1960s. In physics, the Pauli exclusion principle states that two identical subatomic particles, referred to as fermions, cannot occupy the same quantum state simultaneously. This has motivated the use of DPPs to model fermions in thermal equilibrium, but with the name of fermion processes (Macchi, 1975). A probabilistic description of DPPs can be found in Hough et al. (2006) and Borodin (2009).

More recently, DPPs have attracted attention in the machine learning and statistical communities. The work of Kulesza and Taskar (2012) provides a thorough and comprehensible introduction to DPPs. They focused on applications most relevant to the machine learning community, such as search results and document summarization.

An important characteristic of DPPs is that they assign higher probability to sets of items that are diverse. Its name is due to the fact that the likelihood of a DPP depends on the *determinant* of a kernel matrix that defines a global measure of similarity between pairs of items (Borodin and

Olshanski, 2000).

There has been an increasing interest in DPPs in the machine learning community. For instance, Kulesza and Taskar (2011b) focused on structured DPPs when modelling distributions over sets of structures, such as sets of sequences, trees, or graphs; Kulesza and Taskar (2011a) on a maximum *a posteriori* inference-based strategy for learning parameters of a DPP, Affandi et al. (2012) on Markov DPPs, which are advantageous when interest is to sequentially select multiple diverse sets of items. Gillenwater et al. (2012) proposed an approximate optimization solution algorithm for finding the most likely configuration in the DPP modelling framework; Affandi et al. (2013) used a Nyström approximation to project the kernel matrix into a low-dimensional space and improve the feasibility of sampling DPPs in high-dimensional settings. Finally, Affandi et al. (2014) focused on approximate inference for DPPs and the use of Bayesian methods to learn their parameters.

In the statistical literature, Lavancier et al. (2015) focused on the probabilistic properties of DPPs and developed parametric models for analyzing DPPs. Inference was likelihood-based and their methodology was applied to spatial point pattern data sets with the goal of modelling different degrees of repulsiveness. Another recent work in the statistical literature is that of Shirota and Gelfand (2016), which focused on inference in a Bayesian framework via an approximate Bayesian computation approach.

Our motivation is the potential use of these processes in the design of monitoring networks. It is reasonable to assume that if observations are measured in one given location, a designer would not be interested in obtaining measurements at nearby locations of that site, unless there is special interest in the smoothness of the process or in the measurement error. This leads to the idea of spatial repulsion in the design context as it is expected that the study region is to be well covered. We explore the idea that not only a DPP design may be spatially-balanced, but it can also provide a flexible way of imposing diversity in the selection of locations based on additional variables that might be available. We develop these ideas in Chapter 8.

For the purposes of using DPPs in the design of monitoring network

context, we restrict ourselves to introducing DPPs on a discrete finite set $\mathcal{Y} = \{1, \ldots, N\}$, as in Kulesza and Taskar (2012). We believe it is more appropriate for our application as, in practice, the domain in which we are allowed to locate monitoring stations is often discretized, since we are usually restricted by the accessibility for the potentially new sites. Moreover, it will be more computationally efficient to deal with discretized spaces. For some background material on more general DPPs defined on a Borel set $B \subseteq \mathbb{R}^d$, check Lavancier et al. (2015).

The rest of this chapter is as follows. In Section 7.2 we provide some definitions regarding DPPs on a discrete set, as well as an algorithm to sample from it. Finally, in Section 7.3, we describe the notion of k-DPPS, which will be essential for the monitoring of networks context seen in Chapter 8.

7.2 Definitions

A point process $\mathcal{P}$ on a discrete set $\mathcal{Y} = \{1, \ldots, N\}$ is a probability measure on $2^{\mathcal{Y}}$, the set of all possible subsets $Y \subseteq \mathcal{Y}$. Suppose that each element i is included with, say, probability p_i. In an independent point process, the probability of each subset is then

$$\mathcal{P}(Y) = \prod_{i \in Y} p_i \prod_{i \notin Y} (1 - p_i). \tag{7.1}$$

A point process $\mathcal{P}$ is called a determinantal point process if, when $\mathbf{Y}$ is a random subset drawn according to $\mathcal{P}$, for every $A \subseteq \mathcal{Y}$,

$$\mathcal{P}(A \subseteq \mathbf{Y}) = \det(K_A), \tag{7.2}$$

where K is a real, symmetric $N \times N$ positive semidefinite matrix indexed by the elements of $\mathcal{Y}$, and $K_A \equiv [K_{ij}]_{i,j \in A}$, such that $\det(K_\emptyset) = 1$, where $\emptyset$ denotes the empty set.

Note that equation (7.2) defines marginal probabilities of inclusion for subsets A. In particular, if one is interested in the probability of a particular

item i being selected, i.e, when $A = \{i\}$, then

$$\mathcal{P}(i \in \mathbf{Y}) = \det(K_{ii}) = K_{ii}. \tag{7.3}$$

Hence, the diagonal entries of K give the marginal probabilities of inclusion for individual elements.

DPPs are parametrized by this marginal kernel matrix K which defines a global measure of similarity between pairs of items, so that more similar items are less likely to co-occur. Thus, a DPP assigns higher probability to sets of items that are diverse. Note that in the case when $A = \{i, j\}$,

$$\begin{aligned} \mathcal{P}(i, j \in \mathbf{Y}) &= \det \begin{pmatrix} K_{ii} & K_{ij} \\ K_{ji} & K_{jj} \end{pmatrix} \\ &= K_{ii}K_{jj} - K_{ij}K_{ji} \\ &= \mathcal{P}(i \in \mathbf{Y})\mathcal{P}(j \in \mathbf{Y}) - K_{ij}^2, \end{aligned} \tag{7.4}$$

so large values of K_{ij} imply that i and j tend not to co-occur.

In order to fully characterize a DPP, the eigenvalues of the marginal kernel K need to be bounded above by one. Hence, in practice, it is more convenient to characterize DPPs via L-ensembles (Borodin and Rains, 2005; Kulesza and Taskar, 2012), which directly define the probability of observing each subset of $\mathcal{Y}$. An L-ensemble defines a DPP not through the marginal kernel K, but via a real positive semidefinite matrix L, hereby referred to as L-ensemble, indexed by the elements of $\mathcal{Y}$, such that:

$$\mathcal{P}_L(\mathbf{Y} = Y) \propto \det(L_Y). \tag{7.5}$$

Any positive semidefinite matrix defines a DPP. The normalization constant is available in closed form since

$$\sum_{Y \subseteq \mathcal{Y}} \det(L_Y) = \det(L + I), \tag{7.6}$$

where I is the $N \times N$ identity matrix. Thus,

$$\mathcal{P}_L(Y) = \mathcal{P}_L(\mathbf{Y} = Y) = \frac{\det(L_Y)}{\det(L+I)}. \tag{7.7}$$

In contrast to the previous marginal probabilities of inclusion for subsets A, (7.5) directly considers the probability of exactly observing all possible realizations of Y.

That being said, alternative representations of DPPs are done either through the marginal kernel K or the L-ensemble. It is possible to translate between them (Macchi, 1975) as follows

$$K = (L+I)^{-1}L. \tag{7.8}$$

The algorithm to sample from a DPP is based on an orthonormal eigendecomposition of the marginal kernel, which can be obtained through the eigendecomposition of the positive semidefinite matrix L:

$$L = \sum_{i=1}^{N} \lambda_n \mathbf{v}_n \mathbf{v}_n^\top \tag{7.9}$$

and a rescaling of eigenvalues

$$K = \sum_{i=1}^{N} \frac{\lambda_n}{1+\lambda_n} \mathbf{v}_n \mathbf{v}_n^\top. \tag{7.10}$$

In the above, $\{\mathbf{v}_n, \lambda_n\}$ denote the eigenvectors and eigenvalues of L.

Algorithm 3 summarizes how to sample from a DPP, as originally described in Kulesza and Taskar (2012). In the algorithm, the $\mathbf{e}_i$ vector denotes a zero vector with 1 in its i-th entry. Note that the DPP sampling algorithm first entails sampling a subset of eigenvectors of the L-ensemble, where their associated eigenvalues govern their probabilities of selection. Note that the cardinality of the set of selected eigenvectors selected is unknown in advance, which can simply be viewed as sum of N independent Bernoulli random variables.

Algorithm 3 Sampling from a DPP

Input: $\{\mathbf{v}_n, \lambda_n\}$ eigenvectors and eigenvalues of L
  $J \leftarrow \emptyset$
  for $n = 1, \ldots, N$ **do**
    $J \leftarrow J \cup \{n\}$ with probability $\frac{\lambda_n}{1+\lambda_n}$
  end for
  $V \leftarrow \{\mathbf{v}_n\}_{n \in J}$
  $Y \leftarrow \emptyset$
  while $|V| > 0$ **do**
    Select y_i from $\mathcal{Y}$ with probability given by $\frac{1}{|V|} \sum_{\mathbf{v} \in V} (\mathbf{v}^\top \mathbf{e}_i)^2$
    $Y \leftarrow Y \cup y_i$
    $V \leftarrow V_\perp$, an orthonormal basis of the subspace of V orthogonal to $\mathbf{e}_i$
  end while
Output: Y

An approximation to a Poisson process in a plane where points are sampled independently seen in Figure 7.1 is contrasted with a simulation from a DPP using a Gaussian kernel. The DPP simulation displays much less clumping, providing a better coverage of that region.

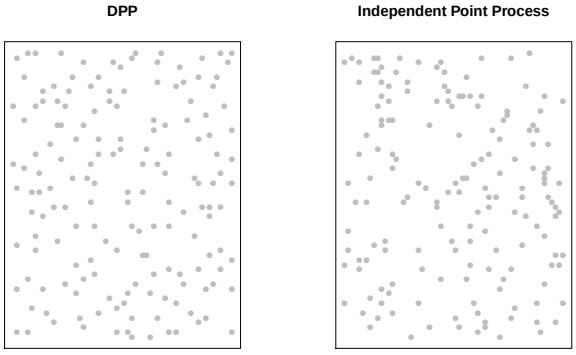


Figure 7.1: A set of points in the plane drawn from (left) a DPP characterized by an L-ensemble with Gaussian kernel and (right) the same number of points sampled independently. Note the clumping associated to the randomly sampled points in contrast to the more spatially balanced set of points sampled from the DPP.

Before we describe DPPs with fixed cardinality, we need to introduce an important way of expressing a DPP as a mixture of elementary DPPs (Kulesza and Taskar, 2012), also commonly known as determinantal projection processes. Elementary DPPs, denoted as $\mathcal{P}^V$, are a particular type of DPP where every eigenvalue of its marginal kernel is either zero or one. Its marginal kernel can thus be decomposed as

$$K^V = \sum_{\mathbf{v} \in V} \mathbf{v}\mathbf{v}^\top, \tag{7.11}$$

where V is a set of orthonormal vectors. From this decomposition, note that elementary DPPs have their cardinality fixed as the cardinality of V. Furthermore, denoting V_J as $\{\mathbf{v}_n\}_{n \in J}$, the mixture is (Kulesza and Taskar, 2012)

$$\mathcal{P}_L(Y) = \frac{1}{\det(L + I)} \sum_{J \subseteq \{1,\ldots,N\}} \mathcal{P}^{V_j}(Y) \prod_{n \in J} \lambda_n. \tag{7.12}$$

The notion of elementary DPPs will be particularly useful to define the normalization constant under fixed cardinality DPPs, which is introduced in the following Section 7.3.

7.3 k-DPPs

A k-DPP on a discrete set $\mathcal{Y} = \{1, \ldots, N\}$ is simply a DPP with fixed cardinality k. The modelling is thus restricted on which elements of size k are part of a random subset of $\mathcal{Y}$. This notion is essential for the monitoring of networks context seen in Chapter 8. In practice, the number of monitoring sites that one can afford to sample from is usually known in advance. On the other hand, standard DPPs models may yield subsets of any size.

A k-DPP can be obtained by conditioning a standard DPP on the event that the set Y has cardinality k, as follows

$$\mathcal{P}_L^k(Y) = \mathcal{P}(\mathbf{Y} = Y \mid |Y| = k) = \frac{\det(L_Y)}{\sum_{|Y'|=k} \det(L_{Y'})}, \tag{7.13}$$

where $|Y|$ denotes the cardinality of Y and L is a positive semidefinite matrix indexed by the elements of $\mathcal{Y}$.

From the notion of elementary DPPs, note that the normalization constant of the k-DPP is given by

$$\begin{aligned}
\sum_{|Y'|=k} \det(L_{Y'}) &= \det(L+I) \sum_{|Y'|=k} \mathcal{P}_L(Y') && (7.14)\\
&= \sum_{|Y'|=k} \sum_{J \subseteq \{1,\ldots,N\}} \mathcal{P}^{V_j}(Y') \prod_{n \in J} \lambda_n && (7.15)\\
&= \sum_{\substack{J \subseteq \{1,\ldots,N\} \\ |J|=k}} \prod_{n \in J} \lambda_n && (7.16)\\
&\equiv E_k^N, && (7.17)
\end{aligned}$$

which can be computed recursively noting that

$$E_k^N = E_k^{N-1} + \lambda_N E_{k-1}^{N-1}, \tag{7.18}$$

where $\lambda_1, \ldots, \lambda_N$ are the eigenvalues of the L-ensemble.

Algorithm 4 summarizes how to sample from a k-DPP. The main difference of the k-DPP algorithm is in its first step, where this time the subset of eigenvectors is sampled with a fixed cardinality k.

Algorithm 4 Sampling from a k-DPP

Input: size k and $\{\mathbf{v}_n, \lambda_n\}$ eigenvectors and eigenvalues of L

$J \leftarrow \emptyset$

Compute $E_1^n, \ldots, E_k^n$, for $n = 0, \ldots, N$

for $n = N, \ldots, 1$ **do**

Sample $u \sim U[0, 1]$

if $u < \frac{\lambda_n E_{k-1}^{n-1}}{E_k^n}$ **then**

$J \leftarrow J \cup \{n\}$

$k \leftarrow k - 1$

if k $= 0$ **then**

break

end if

end if

end for

$V \leftarrow \{\mathbf{v}_n\}_{n \in J}$

$Y \leftarrow \emptyset$

while $|V| > 0$ **do**

Select y_i from $\mathcal{Y}$ with probability given by $\frac{1}{|V|} \sum_{\mathbf{v} \in V} (\mathbf{v}^\top \mathbf{e}_i)^2$

$Y \leftarrow Y \cup \{y_i\}$

$V \leftarrow V_\perp$, an orthonormal basis for the subspace of V orthogonal to $\mathbf{e}_i$

end while

Output: Y

Chapter 8

Design of Monitoring Networks

8.1 Importance of Designing Monitoring Networks

From an environmental perspective, monitoring networks play an important role in surveillance of environmental processes which may impact either human health or nature. The measurements obtained from the monitoring of temperature, precipitation, and pollutants within a region, for instance, provide critical data for both scientists and governmental agencies that can be used for many essential objectives, such as

- Are the measurements obtained above the regulatory limits?
- Is there a trend in a given health outcome that could potentially be associated with an environmental hazard?

With the advances in geographic information systems (GIS) (Goodchild and Haining, 2004; Murray, 2010), modern agriculture now relies on interactive tools that provide climate and soil information. Such information is valuable not only to maximize yield but also for the development of more environmentally friendly practices.

However, with important objectives come interesting challenges. Many design strategies have been developed. The different strategies are based on providing a good spatial coverage of the domain, by ensuring randomization and selecting locations at random given a certain probability of inclusion; or even depending on a model used to learn about a particular underlying

phenomenon. A brief description of such methods can be found in Section 8.3.

Moreover, it is important to note that a monitoring network need not be static. Not only may a network's purpose change over time, regulatory budgets may also allow new sites to be added or may impose a reduction in the network.

Most importantly, we would like to advocate the need for "mobile friendly" design strategies, i.e. approaches suitable for the design of dynamic or mobile networks. In the agricultural context, weather conditions or other sudden emerging risks may require a close surveillance of the agriculture fields, and their design purpose may quickly and dramatically change over time. The surveillance itself can be done by obtaining measurements at certain key locations in the field using portable devices (Tothill, 2001; Rodriguez-Mozaz et al., 2006), such as nutrient meters for obtaining soil macro-nutrient information. Depending on how these new conditions are expected to impact their crop, there may be a need to dynamically choose new locations on where to measure.

In practice, however, stations are often preferentially deployed, possibly restricted to the accessibility of the potential new sites and budgetary constraints. Diggle et al. (2010) drew attention to inference under this scenario. Preferential sampling occurs when the sampling locations are stochastically dependent on the underlying process. A clear example is in air pollution monitoring. Data are sometimes collected at locations where it is anticipated that the outcome will have a large or small value (Guttorp and Sampson, 2010).

Diggle et al. (2010) pointed out that ignoring preferential sampling in geostatistical models can lead to misleading inferences. In order to adjust for potential biases, they proposed a shared latent process model for geostatistical modelling with preferentially sampled data. This issue is similarly discussed in the Bayesian framework by Pati et al. (2011). Gelfand et al. (2012) suggested a simulation-based approach to assess the effects of preferential sampling based on information about underlying process and other factors known drive that process. Also, Zidek et al. (2014) suggested a bias

correction approach that is able to accommodate changes to the network due to preferential sampling over time. Ferreira and Gamerman (2015) use an approach based on utility functions in order to analyze the influence of preferential sampling in situations where the goal is to optimize an objective function.

The cited recent works have taken a negative view towards preferential sampling. Though it is clear that the effects of preferential sampling on both estimation and prediction should not be disregarded, we explore how we can obtain balanced designs yielding a high-quality yet diverse set of monitors. We introduce a flexible design strategy based on k-DPPs (Section 7.3) that is able to yield a spatially balanced design by imposing repulsion on the distances between the candidate locations and hence avoid spatial clumping, but it also has the ability to assess similarity between the potential locations should there be extra sources of information related to the underlying process of interest.

8.2 Contributions

In this chapter, our main contribution is the introduction of a novel flexible monitoring network design strategy based on k-DPPs. This strategy is able to handle both designing and redesigning a monitoring network. The k-DPP design can yield spatially balanced designs by imposing repulsion on the distances between the candidate locations. It is also possible to assess similarity and impose repulsion between the potential locations based on extra sources of information that might be related to the phenomenon of interest.

Since its essence is that of a randomized design, we explore its potential use as a randomized alternative to space-filling designs. An overview of space-filling designs can be found in Section 8.3.1.

Additionally, the k-DPP optimal design objective is remarkably similar to that of entropy design, which is reviewed in Section 8.3.2. Since the optimization for entropy designs is a NP-hard problem (Ko et al., 1995), we explore a sampling design for approximating the entropy design optimal so-

lution. This strategy explores the stochasticity of k-DPP and the simplicity of obtaining samples from this process.

8.3 A Review of Design Strategies

Design strategies are often divided into the following groups (Zidek and Zimmerman, 2010):

- **Geometry-based designs**: Designs of this type are based solely on geometric considerations and, in general, their intent is to provide a good coverage of the design region. An example is the space-filling design (Cox et al., 1997; Nychka et al., 1997; Royle and Nychka, 1998; Van Groenigen et al., 2000), which we describe in Section 8.3.1. These designs are particularly useful for exploratory purposes (Müller, 2005). However, as a non-randomized design strategy, it may be prone to potential sampling biases.

- **Probability-based designs**: These designs are often based on randomly sampling locations from the design region. Although they can avoid potential sampling biases due to the randomization, the sampled points may be clumped in some particular areas of the design region. Considering that nearby locations tend to share similar characteristics, sampling locations too close to each other may not bring valuable information about the process of interest.

- **Model-based designs**: As pointed out by Zidek and Zimmerman (2010), environmental monitoring networks are usually based on model-based designs. These designs often optimize a certain characteristic about the process of interest, such as the reduction of uncertainty about model parameters or of the uncertainty about the prediction at unmeasured locations.

In the following subsections, we briefly expand on the descriptions for space-filling and entropy-based designs.

8.3.1 Space-Filling Designs

Space-filling (SF) spatial designs aim at providing a good coverage of the design region based on a criterion that is purely geometric-based. SF designs thus make use of geometric measures to assess the coverage quality. Outside of the monitoring network context, SF designs have also been quite extensively explored in computer experiments as a way of selecting inputs (Sacks et al., 1989; Haaland et al., 1994). A survey of SF designs for computer experiments can be found in Pronzato and Müller (2012).

Denote by $\mathcal{C}$ a set of candidate points, usually based on a fine discretization of the design region, and let $\mathcal{D} \subset \mathcal{C}$ denote the set of k design points. A metric for the distance of any point $\mathbf{s}$ and a particular design $\mathcal{D}$ is given by (Nychka et al., 1997; Royle and Nychka, 1998):

$$d_p(\mathbf{s}, \mathcal{D}) = \left(\sum_{\mathbf{u} \in \mathcal{D}} ||\mathbf{s} - \mathbf{u}||^p \right)^{1/p}. \tag{8.1}$$

This metric determines how well the design covers the point $\mathbf{s}$. Their overall coverage criterion is based on minimizing averages of the coverage metric for every candidate point, given by

$$C_{p,q}(\mathcal{D}) = \left(\sum_{\mathbf{u} \in \mathcal{C}} d_p(\mathbf{s}, \mathcal{D})^q \right)^{1/q}, \tag{8.2}$$

for all $\mathcal{D} \subset \mathcal{C}$. Royle and Nychka (1998) describe a suboptimal solution based on a point swapping algorithm. The algorithm uses random starting configurations and aim at decreasing the coverage criterion by swapping candidate and design points until convergence.

An advantage of this method is its computational simplicity, which is currently implemented in the `R` package `fields` (Nychka et al., 2015). Royle and Nychka (1998) point out that the resulting designs are nearly optimal for spatial prediction purposes.

A generalization of the space-filling design is that of Van Groenigen et al. (2000) who propose a weighted means of shortest distances criterion based

on minimizing

$$\int_A d(\mathbf{s})w(\mathbf{s})d\mathbf{s}, \tag{8.3}$$

where $A \subseteq \mathbb{R}^2$, $w(\mathbf{s})$ is a weight function, and the distance between $\mathbf{s} \in A$ and its closest design points is given by

$$d(\mathbf{s}) = \min_i ||\mathbf{s} - \mathbf{x}_i||. \tag{8.4}$$

However, since these design strategies are geometry-based, they do not allow inclusion of other sources of information, and are highly dependent on the coverage criterion. As non-randomized design strategies, they may be prone to potential sampling biases.

8.3.2 Entropy-Based Designs

The uncertainty about $\mathbf{Y}$ can be represented by the the entropy of its distribution

$$H(\mathbf{Y}) = \mathbb{E}_{\mathbf{Y}}\left[-\log\left(\frac{f(\mathbf{Y})}{h(\mathbf{Y})}\right)\right], \tag{8.5}$$

where $h(\mathbf{Y})$ denotes a reference density, which need not be integrable, allowing the entropy to be invariant under one-to-one transformations of the scale of $\mathbf{Y}$ (Jaynes, 1963).

In hierarchical models for environmental processes, $\mathbf{Y}$ is usually defined conditionally on some hyperparameters, which we denote by $\mathbf{\Psi}$. Considering minimizing uncertainty about $\mathbf{\Psi}$ as another design objective (Caselton et al., 1992), the total entropy can be defined as

$$
\begin{aligned}
H(\mathbf{Y}, \boldsymbol{\Psi}) &= \mathbb{E}\left[-\log\left(\frac{f(\mathbf{Y}, \boldsymbol{\Psi})}{h_{\mathbf{Y},\boldsymbol{\Psi}}(\mathbf{Y}, \boldsymbol{\Psi})}\right)\right] && (8.6)\\
&= \mathbb{E}\left[-\log\left(\frac{f(\mathbf{Y}|\boldsymbol{\Psi})f(\boldsymbol{\Psi})}{h_{\mathbf{Y}}(\mathbf{Y})h_{\boldsymbol{\Psi}}(\boldsymbol{\Psi})}\right)\right] && (8.7)\\
&= \mathbb{E}\left[-\log\left(\frac{f(\mathbf{Y}|\boldsymbol{\Psi})}{h_{\mathbf{Y}}(\mathbf{Y})}\right)\right] + \mathbb{E}\left[-\log\left(\frac{f(\boldsymbol{\Psi})}{h_{\boldsymbol{\Psi}}(\boldsymbol{\Psi})}\right)\right] && (8.8)\\
&= H(\mathbf{Y}|\boldsymbol{\Psi}) + H(\boldsymbol{\Psi}). && (8.9)
\end{aligned}
$$

Similarly, we can define the total entropy in the context of the design of monitoring networks by first augmenting the data into $\mathbf{Y} = (\mathbf{Y}^{(u)}, \mathbf{Y}^{(g)})$, where $\mathbf{Y}^{(u)}$ denotes the measurements at potential sites, currently ungauged, and $\mathbf{Y}^{(g)}$ relates to the existing sites, referred to as gauged locations. The result is:

$$
\begin{aligned}
H(\mathbf{Y}^{(u)}, \mathbf{Y}^{(g)}, \boldsymbol{\Psi}) &= \mathbb{E}\left[-\log\left(\frac{f(\mathbf{Y}^{(u)}, \mathbf{Y}^{(g)}, \boldsymbol{\Psi})}{h_{\mathbf{Y}^{(u)},\mathbf{Y}^{(g)},\boldsymbol{\Psi}}(\mathbf{Y}^{(u)}, \mathbf{Y}^{(g)}, \boldsymbol{\Psi})}\right)\right]\\
&= \mathbb{E}\left[-\log\left(\frac{f(\mathbf{Y}^{(u)}|\mathbf{Y}^{(g)}, \boldsymbol{\Psi}) \times f(\boldsymbol{\Psi}|\mathbf{Y}^{(g)}) \times f(\mathbf{Y}^{(g)})}{h_{\mathbf{Y}^{(u)}}(\mathbf{Y}^{(u)}) \times h_{\boldsymbol{\Psi}}(\boldsymbol{\Psi}) \times h_{\mathbf{Y}^{(g)}}(\mathbf{Y}^{(g)})}\right)\right]\\
&= \mathbb{E}\left[-\log\left(\frac{f(\mathbf{Y}^{(u)}|\mathbf{Y}^{(g)}, \boldsymbol{\Psi})}{h_{\mathbf{Y}^{(u)}}(\mathbf{Y}^{(u)})}\right)\right]\\
&+ \mathbb{E}\left[-\log\left(\frac{f(\boldsymbol{\Psi}|\mathbf{Y}^{(g)})}{h_{\boldsymbol{\Psi}}(\boldsymbol{\Psi})}\right)\right] + \mathbb{E}\left[-\log\left(\frac{f(\mathbf{Y}^{(g)})}{h_{\mathbf{Y}^{(g)}}(\mathbf{Y}^{(g)})}\right)\right]\\
&= \underbrace{H(\mathbf{Y}^{(u)}|\mathbf{Y}^{(g)}, \boldsymbol{\Psi}) + H(\boldsymbol{\Psi}|\mathbf{Y}^{(g)})}_{H(\mathbf{Y}^{(u)},\boldsymbol{\Psi}|\mathbf{Y}^{(g)})} + H(\mathbf{Y}^{(g)}). && (8.10)
\end{aligned}
$$

The design criterion is based on minimizing $H(\mathbf{Y}^{(u)}, \boldsymbol{\Psi}|\mathbf{Y}^{(g)})$, which measures the uncertainty about $\mathbf{Y}^{(u)}$ and $\boldsymbol{\Psi}$ after $\mathbf{Y}^{(g)}$ is observed. Since the total entropy $H(\mathbf{Y}^{(u)}, \mathbf{Y}^{(g)}, \boldsymbol{\Psi})$ is fixed, an equivalent criterion is to maximize $H(\mathbf{Y}^{(g)})$. Moreover, the same criterion of maximizing $H(\mathbf{Y}^{(g)})$ would be similarly obtained had we decomposed $H(\mathbf{Y}^{(u)}, \mathbf{Y}^{(g)})$ instead of $H(\mathbf{Y}^{(u)}, \mathbf{Y}^{(g)}, \boldsymbol{\Psi})$.

Entropy of Multivariate Normal Distributions

Recall that the log-density of a p-dimensional multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ is given by

$$\log f(\mathbf{Y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{p}{2}\log(2\pi) - \frac{1}{2}\log\det(\boldsymbol{\Sigma}) - \frac{1}{2}(\mathbf{Y}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{Y}-\boldsymbol{\mu}),$$

where $\det(\cdot)$ denotes matrix determinant. Without loss of generality, assuming a reference measure $h_{\mathbf{Y}}(\mathbf{Y}) = 1$, then

$$\begin{aligned} H(\mathbf{Y}) &= \mathbb{E}_{\mathbf{Y}}\left[\frac{p}{2}\log(2\pi) + \frac{1}{2}\log\det(\boldsymbol{\Sigma}) + \frac{1}{2}(\mathbf{Y}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{Y}-\boldsymbol{\mu})\right] \\ &= \frac{p}{2}\log(2\pi) + \frac{1}{2}\log\det(\boldsymbol{\Sigma}) + \frac{1}{2}\mathbb{E}_{\mathbf{Y}}[(\mathbf{Y}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{Y}-\boldsymbol{\mu})] \\ &= \frac{p}{2}\log(2\pi) + \frac{1}{2}\log\det(\boldsymbol{\Sigma}) + \frac{p}{2} \\ &= \frac{p}{2}[\log(2\pi)+1] + \frac{1}{2}\log\det(\boldsymbol{\Sigma}). \end{aligned} \tag{8.11}$$

Now suppose that we can partition $\mathbf{Y}$ as $(\mathbf{Y}^{(u)}, \mathbf{Y}^{(g)})$. Denoting the matrices $\boldsymbol{\Sigma}_{gg}$ and $\boldsymbol{\Sigma}_{uu}$ as the covariance matrices of $\mathbf{Y}^{(g)}$ and $\mathbf{Y}^{(u)}$, respectively, and $\boldsymbol{\Sigma}_{ug}$ the cross-covariance, we can partition $\boldsymbol{\Sigma}$ as

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{uu} & \boldsymbol{\Sigma}_{ug} \\ \boldsymbol{\Sigma}_{gu} & \boldsymbol{\Sigma}_{gg} \end{pmatrix}. \tag{8.12}$$

Recall from the properties of the multivariate normal that $\mathbf{Y}^{(u)}|\mathbf{Y}^{(g)} \sim \mathcal{N}(\boldsymbol{\mu}_{u|g}, \boldsymbol{\Sigma}_{u|g})$, where $\boldsymbol{\Sigma}_{u|g} = \boldsymbol{\Sigma}_{uu} - \boldsymbol{\Sigma}_{ug}\boldsymbol{\Sigma}_{gg}^{-1}\boldsymbol{\Sigma}_{gu}$. Hence, the entropy of $\mathbf{Y}$ can be written as

$$H(\mathbf{Y}^{(u)}, \mathbf{Y}^{(g)}) = H(\mathbf{Y}^{(u)}|\mathbf{Y}^{(g)}) + H(\mathbf{Y}^{(g)}) \tag{8.13}$$

$$\overset{c}{\propto} \frac{1}{2}\log\det(\boldsymbol{\Sigma}_{u|g}) + \frac{1}{2}\log\det(\boldsymbol{\Sigma}_{gg}), \tag{8.14}$$

where $\overset{c}{\propto}$ denotes proportionality up to additive constants. The entropy criterion is thus to minimize $\log\det(\boldsymbol{\Sigma}_{u|g})$, or equivalently, to maximize $\log\det(\boldsymbol{\Sigma}_{gg})$.

In the context of monitoring networks, say, when the goal is to augment the network, the objective is to find a subset of u^+ sites among the u ungauged ones (also referred to as candidate sites, where $\mathcal{C}$ denote the set of candidate points) to add to the existing network. We denote the remaining sites that are not the selected as u^-. The resulting network will then consist of $(\mathbf{Y}^{(u^+)}, \mathbf{Y}^{(g)})$.

Note that $\mathbf{Y}^{(u)}$ can be partitioned into $(\mathbf{Y}^{(u^+)}, \mathbf{Y}^{(u^-)})$. Proceeding similarly as above, note that

$$\begin{aligned} H(\mathbf{Y}^{(u^+)}, \mathbf{Y}^{(u^-)}, \mathbf{Y}^{(g)}) &= H(\mathbf{Y}^{(u^+)}, \mathbf{Y}^{(u^-)} | \mathbf{Y}^{(g)}) + H(\mathbf{Y}^{(g)}) \\ &= H(\mathbf{Y}^{(u^-)} | \mathbf{Y}^{(u^+)}, \mathbf{Y}^{(g)}) + H(\mathbf{Y}^{(u^+)}, \mathbf{Y}^{(g)}). \end{aligned}$$

Since the total entropy $H(\mathbf{Y}^{(u^+)}, \mathbf{Y}^{(u^-)}, \mathbf{Y}^{(g)})$ is fixed, it will be optimal to augment the network with the u^+ sites so as to maximize $H(\mathbf{Y}^{(u^+)}, \mathbf{Y}^{(g)})$.

Considering that

$$\begin{aligned} H(\mathbf{Y}^{(u^+)}, \mathbf{Y}^{(g)}) &= H(\mathbf{Y}^{(u^+)} | \mathbf{Y}^{(g)}) + H(\mathbf{Y}^{(g)}) && (8.15) \\ &\overset{c}{\propto} \frac{1}{2} \log \det(\mathbf{\Sigma}_{u^+|g}) + \frac{1}{2} \log \det(\mathbf{\Sigma}_{gg}), && (8.16) \end{aligned}$$

it will be optimal to maximize $\frac{1}{2} \log \det(\mathbf{\Sigma}_{u^+|g})$. The entropy criterion for augmenting the network is thus

$$\arg \max_{u^+ \subset \mathcal{C}} \frac{1}{2} \log |\mathbf{\Sigma}_{u|g}|. \tag{8.17}$$

Entropy of Multivariate t-Distributions

The entropy of a multivariate t-distribution can be obtained as a scale mixture of a multivariate normal and an inverted Wishart distribution (Caselton et al., 1992; Guttorp et al., 1993). Let $\mathbf{Y}$ be a g-dimensional random variable such as

$$\begin{aligned} \mathbf{Y}|\mathbf{\Sigma} &\sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma}) && (8.18) \\ \mathbf{\Sigma}|\mathbf{\Xi}, \delta &\sim \mathcal{W}^{-1}(\mathbf{\Xi}, \delta). && (8.19) \end{aligned}$$

Hence,

$$\mathbf{Y} \sim t\left(\boldsymbol{\mu}, \frac{\boldsymbol{\Xi}}{\delta - g + 1}, \delta - g + 1\right). \tag{8.20}$$

Conditionally on the hyperparameters $\boldsymbol{\Xi}$ and δ, note that the entropy of $\mathbf{Y}$ is defined as

$$H(\mathbf{Y}) = H(\mathbf{Y}|\boldsymbol{\Sigma}) + H(\boldsymbol{\Sigma}) - H(\boldsymbol{\Sigma}|\mathbf{Y}), \tag{8.21}$$

since the joint entropy $H(\mathbf{Y}, \boldsymbol{\Sigma})$ can be decomposed in the following two different ways:

$$\begin{aligned} H(\mathbf{Y}, \boldsymbol{\Sigma}) &= H(\mathbf{Y}|\boldsymbol{\Sigma}) + H(\boldsymbol{\Sigma}) & (8.22)\\ H(\mathbf{Y}, \boldsymbol{\Sigma}) &= H(\boldsymbol{\Sigma}|\mathbf{Y}) + H(\mathbf{Y}). & (8.23) \end{aligned}$$

Assuming the following reference measure $h(\mathbf{Y}, \boldsymbol{\Sigma}) = |\boldsymbol{\Sigma}|^{-(g+1)/2}$ as in Caselton et al. (1992), we will now describe each component of the entropy of $\mathbf{Y}$.

Since $\mathbf{Y}|\boldsymbol{\Sigma}$ is multivariate normal, using results from Section 8.3.2, and since $\boldsymbol{\Xi}\boldsymbol{\Sigma}^{-1} \sim \mathcal{W}(\mathbf{I}, \delta)$, then

$$\begin{aligned} H(\mathbf{Y}|\boldsymbol{\Sigma}) &= \frac{1}{2}[\log(2\pi) + 1] + \frac{1}{2}\mathbb{E}[\log\det(\boldsymbol{\Sigma}) \mid \boldsymbol{\Xi}] \\ &= \frac{1}{2}[\log(2\pi) + 1] + \frac{1}{2}\mathbb{E}[\log\det(\boldsymbol{\Sigma}\boldsymbol{\Xi}^{-1}\boldsymbol{\Xi}) \mid \boldsymbol{\Xi}] \\ &= \frac{1}{2}[\log(2\pi) + 1] + \frac{1}{2}\mathbb{E}[\log\det(\boldsymbol{\Sigma}\boldsymbol{\Xi}^{-1}) \mid \boldsymbol{\Xi}] + \frac{1}{2}\log\det(\boldsymbol{\Xi}) \\ &\overset{c}{\propto} \frac{1}{2}\log\det(\boldsymbol{\Xi}). \end{aligned} \tag{8.24}$$

Now notice that the other two components of the entropy of $\mathbf{Y}$ in (8.21) are in fact constants. Recall that if $\boldsymbol{\Sigma}|\boldsymbol{\Xi}, \delta \sim \mathcal{W}^{-1}(\boldsymbol{\Xi}, \delta)$ then its density can be written as

$$f(\boldsymbol{\Sigma}) \propto \det(\boldsymbol{\Xi})^{\delta/2} \det(\boldsymbol{\Sigma})^{-\frac{\delta+p+1}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Xi}\boldsymbol{\Sigma}^{-1})\right\}. \tag{8.25}$$

Due to our choice of reference measure, conditionally on the hyperparameters $\boldsymbol{\Xi}$ and δ, the entropy of $\boldsymbol{\Sigma}$ is

$$\begin{aligned} H(\boldsymbol{\Sigma}) &= \mathbb{E}\left[-\log\left(\frac{f(\boldsymbol{\Sigma})}{h(\boldsymbol{\Sigma})}\right)\right] \\ &\overset{c}{\propto} -\frac{\delta}{2}\log\det(\boldsymbol{\Xi}) + \frac{\delta}{2}\mathbb{E}(\log\det(\boldsymbol{\Sigma}) \mid \boldsymbol{\Xi}) + \frac{1}{2}\mathbb{E}[\mathrm{tr}(\boldsymbol{\Xi}\boldsymbol{\Sigma}^{-1}) \mid \boldsymbol{\Xi}] \\ &= -\frac{\delta}{2}\log\det(\boldsymbol{\Xi}) + \frac{\delta}{2}\mathbb{E}[\log\det(\boldsymbol{\Sigma}\boldsymbol{\Xi}^{-1}\boldsymbol{\Xi}) \mid \boldsymbol{\Xi}] + \frac{1}{2}\mathbb{E}[\mathrm{tr}(\boldsymbol{\Xi}\boldsymbol{\Sigma}^{-1}) \mid \boldsymbol{\Xi}] \\ &= -\frac{\delta}{2}\log\det(\boldsymbol{\Xi}) + \frac{\delta}{2}\mathbb{E}[\log\det(\boldsymbol{\Sigma}\boldsymbol{\Xi}^{-1}) \mid \boldsymbol{\Xi}] + \frac{\delta}{2}\log\det(\boldsymbol{\Xi}) \\ &\quad + \frac{1}{2}\mathbb{E}[\mathrm{tr}(\boldsymbol{\Xi}\boldsymbol{\Sigma}^{-1}) \mid \boldsymbol{\Xi}] \\ &\overset{c}{\propto} c_1(p, \delta), \end{aligned} \tag{8.26}$$

where $c_1(p, \delta)$ denotes a constant that depends on p and δ. This is due to the fact that $\boldsymbol{\Xi}\boldsymbol{\Sigma}^{-1} \sim \mathcal{W}(\mathbf{I}, \delta)$.

Similarly, conditionally on the hyperparameters $\boldsymbol{\Xi}$ and δ, we will derive the entropy of $\boldsymbol{\Sigma}|\mathbf{Y}$. Firstly, note that the marginal posterior of $\boldsymbol{\Sigma}$ is given by

$$\boldsymbol{\Sigma}|\mathbf{Y} \sim \mathcal{W}^{-1}(\boldsymbol{\Xi} + \mathbf{Y}\mathbf{Y}^\top, \delta + 1). \tag{8.27}$$

Recall that the reference measure we are using is $h(\mathbf{Y}, \boldsymbol{\Sigma}) = h(\mathbf{Y})h(\boldsymbol{\Sigma}) = |\boldsymbol{\Sigma}|^{-(g+1)/2}$. Hence, conditionally on the hyperparameters $\boldsymbol{\Xi}$ and δ, the

entropy of $\boldsymbol{\Sigma}|\mathbf{Y}$ is given by

$$\begin{aligned}
H(\boldsymbol{\Sigma} \mid \mathbf{Y}) &= \mathbb{E}\left[-\log\left(\frac{f(\boldsymbol{\Sigma}|\mathbf{Y})}{h(\boldsymbol{\Sigma})}\right)\right] \\
&\overset{c}{\propto} -\frac{\delta+1}{2}\mathbb{E}[\log\det(\boldsymbol{\Xi}+\mathbf{Y}\mathbf{Y}^\top) \mid \boldsymbol{\Xi}] \\
&\quad +\frac{\delta+1}{2}\mathbb{E}[\log\det(\boldsymbol{\Sigma}) \mid \boldsymbol{\Xi}] + \frac{1}{2}\mathbb{E}[\mathrm{tr}((\boldsymbol{\Xi}+\mathbf{Y}\mathbf{Y}^\top)\boldsymbol{\Sigma}^{-1}) \mid \boldsymbol{\Xi}] \\
&= -\frac{\delta+1}{2}\log\det(\boldsymbol{\Xi}) - \frac{\delta+1}{2}\mathbb{E}[\log(\mathbf{Y}^\top\boldsymbol{\Xi}^{-1}\mathbf{Y}) \mid \boldsymbol{\Xi}] \\
&\quad +\frac{\delta+1}{2}\mathbb{E}[\log\det(\boldsymbol{\Sigma}\boldsymbol{\Xi}^{-1}) \mid \boldsymbol{\Xi}] + \frac{\delta+1}{2}\log\det(\boldsymbol{\Xi}) \\
&\quad +\frac{1}{2}\mathbb{E}[\mathrm{tr}((\boldsymbol{\Xi}+\mathbf{Y}\mathbf{Y}^\top)\boldsymbol{\Sigma}^{-1}) \mid \boldsymbol{\Xi}] \\
&\overset{c}{\propto} -\frac{\delta+1}{2}\mathbb{E}[\log(\mathbf{Y}^\top\boldsymbol{\Xi}^{-1}\mathbf{Y}) \mid \boldsymbol{\Xi}] + \frac{\delta+1}{2}\mathbb{E}[\log\det(\boldsymbol{\Sigma}\boldsymbol{\Xi}^{-1}) \mid \boldsymbol{\Xi}] \\
&\quad +\frac{1}{2}\mathbb{E}[\mathrm{tr}(\boldsymbol{\Xi}\boldsymbol{\Sigma}^{-1}) \mid \boldsymbol{\Xi}] + \frac{1}{2}\mathbb{E}[\mathrm{tr}(\mathbf{Y}^\top\boldsymbol{\Sigma}^{-1}\mathbf{Y}) \mid \boldsymbol{\Xi}] \\
&\overset{c}{\propto} c_2(p,\delta), \qquad (8.28)
\end{aligned}$$

where $c_2(p,\delta)$ denotes a constant that depends on p and δ. This is due to the fact that

$$\begin{aligned}
\det(\boldsymbol{\Xi}+\mathbf{Y}\mathbf{Y}^\top) &= \det(\boldsymbol{\Xi})[\mathbf{Y}^\top\boldsymbol{\Xi}^{-1}\mathbf{Y}] \qquad (8.29) \\
\mathrm{tr}((\boldsymbol{\Xi}+\mathbf{Y}\mathbf{Y}^\top)\boldsymbol{\Sigma}^{-1}) &= \mathrm{tr}(\boldsymbol{\Xi}\boldsymbol{\Sigma}^{-1}) + \mathrm{tr}(\mathbf{Y}\boldsymbol{\Sigma}^{-1}\mathbf{Y}^\top), \qquad (8.30)
\end{aligned}$$

$\boldsymbol{\Xi}\boldsymbol{\Sigma}^{-1} \sim \mathcal{W}(\mathbf{I},\delta)$, and that given $\boldsymbol{\Xi}$, $\mathbf{Y}^\top\boldsymbol{\Xi}^{-1}\mathbf{Y}$ follows a F-distribution with degrees of freedom depending on p and δ.

Finally, combining (8.24), (8.26) and (8.28), we conclude that the entropy for a $t\left(\boldsymbol{\mu}, \frac{\boldsymbol{\Xi}}{\delta-g+1}, \delta-g+1\right)$ is

$$H(\mathbf{Y}) \overset{c}{\propto} \frac{1}{2}\log\det(\boldsymbol{\Xi}). \qquad (8.31)$$

8.4 k-DPP Design

We propose a flexible monitoring network design strategy based on a k-DPP. We use the fact that not only could a k-DPP design be spatially-balanced, but could also provide a flexible way of imposing diversity in the selection of locations based on additional variables that might be available. The methodology depends on the well-established theory of k-DPPs, though here it is described in a design of monitoring network setting.

Definition 8.5. *A k-DPP design is characterized by an L-ensemble, denoted by L, i.e. any positive semidefinite matrix indexed by a set of n candidate locations, such that $n \geq k$. The design objective is based on the optimal configuration under a cardinality constraint, as follows*

$$\underset{Y \subseteq \mathcal{Y},\ |Y|=k}{\arg\max}\ \det(L_Y), \tag{8.32}$$

where $\mathcal{Y} = \{1, \ldots, n\}$ and $L_Y \equiv [L_{ij}]_{i,j \in Y}$.

Note that the methodology easily accommodates the objective of reducing the number of sites in a monitoring network. In this case, the elements of L should be indexed by the n set of existing monitoring locations. If the goal is to select k stations for reduction, then the cardinality constraint in the design objective should be the cardinality of the complementary set, $n - k$, as follows

$$\underset{Y \subseteq \mathcal{Y},\ |Y|=n-k}{\arg\max}\ \det(L_Y), \tag{8.33}$$

where $\mathcal{Y} = \{1, \ldots, n\}$. This allows for the selection of the optimum $n - k$ set of monitors that will remain in the network, while the other k will be shut down.

Furthermore, the methodology can also be adapted for use when the goal is to augment the network. Let $\mathcal{C}$ be the set of m potential new locations and $\mathcal{G}$ the set of n existing monitors. Let $\mathcal{Y}$ index the elements of the monitoring network which includes all the n existing monitors as well as the m candidate monitors.

Note that the probabilities of selection need to be described conditionally on the existing monitors. We describe these conditional probabilities as follows.

$$\begin{aligned} \mathcal{P}_L(\mathbf{Y} = \mathcal{G} \cup \mathcal{C} | \mathcal{G} \subseteq \mathbf{Y}) &= \frac{\mathcal{P}_L(\mathbf{Y} = \mathcal{G} \cup \mathcal{C})}{\mathcal{P}_L(\mathcal{G} \subseteq \mathbf{Y})} && (8.34) \\ &\propto \mathcal{P}_L(\mathbf{Y} = \mathcal{G} \cup \mathcal{C}) && (8.35) \\ &\propto \det(L_{\mathcal{C}}^{\mathcal{G}}), && (8.36) \end{aligned}$$

where $L^{\mathcal{G}}$ is an L-ensemble indexed by the elements of $\mathcal{Y} - \mathcal{G}$, and can be obtained as follows (Borodin and Rains, 2005; Kulesza and Taskar, 2012)

$$L^{\mathcal{G}} = [(L + I_{\bar{\mathcal{G}}})^{-1}]_{\bar{\mathcal{G}}} - I, \tag{8.37}$$

where $I_{\bar{\mathcal{G}}}$ is the matrix with ones in the diagonal entries of the elements of $\mathcal{Y} - \mathcal{G}$, and L is an L-ensemble indexed by the elements of $\mathcal{Y}$. Recall that the subscripts denote matrix restriction. For instance, $L_{\mathcal{C}}^{\mathcal{G}}$ is simply the restricted $L^{\mathcal{G}}$, i.e. selecting the rows and columns associated with the elements of candidate set $\mathcal{C}$. Denoting the set of u selected monitors as $\mathbf{U}$, the augmentation design objective can be then described as

$$\underset{\mathbf{U} \subseteq \mathcal{C},\ |\mathbf{U}| = m}{\arg\max} \det(L_{\mathbf{U}}^{\mathcal{G}}). \tag{8.38}$$

One of the limitations of the k-DPP design is that the described optimization problems are NP-hard (Ko et al., 1995; Kulesza and Taskar, 2012). In the following Section 8.5.1, we describe a sampling strategy tool based on the k-DPP design that can also be used to select a subset of points among a candidate set of points, which could ultimately be used to obtain a suboptimal approximation.

8.5.1 k-DPP Sampling Design Strategy

In order to handle the NP-hard optimization problem, we propose a sampling design strategy based on k-DPPs. This sampling design strategy takes

advantage of the ease of sampling from a k-DPP as described in Algorithm 4.

Definition 8.6. *A k-DPP sampling design is a probability-based design that entails sampling from a k-DPP characterized by an L-ensemble, denoted by L, i.e. any positive semidefinite matrix indexed by a set of n candidate locations. Every subset of k locations among the candidate points has the opportunity to be sampled with probability*

$$\mathcal{P}_L^k(Y) = \frac{\det(L_Y)}{\sum_{|Y'|=k} \det(L_{Y'})}.$$

A remarkable characteristic of a k-DPP sampling design is that its methodology is flexible. The design strategies described are governed by the choice of the L-ensemble, which brings some flexibility for the scope of allowable design strategies.

Spatially-balanced k-DPP Designs

A spatially-balanced k-DPP design can be easily constructed as an alternative to space-filling designs when the only information available are the locations of the potential new monitors.

Result 8.6.1. *A spatially-balanced k-DPP design can be used as an alternative to space-filling designs when the only information available are the locations of the potential new monitors. It suffices to construct a positive semidefinite matrix L whose entries depend on a measure of distance between the candidate locations.*

A simple way to construct an L-ensemble of such type is by using a Gaussian kernel, as described below. Note that the positivity of the kernel guarantees that L is indeed a positive semidefinite matrix.

- **Gaussian radial basis kernel**: The (i, j) entries of the L are given by $L_{i,j} = \exp\{-||\mathbf{s}_i - \mathbf{s}_j||^2/2\sigma^2\}$, where $||\mathbf{s}_i - \mathbf{s}_j||$ denotes the Euclidean distance between locations $\mathbf{s}_i$ and $\mathbf{s}_j$.

In Section 8.7, we illustrate how spatially-balanced k-DPP designs can be used as an alternative to space-filling designs.

Inclusion of Extra Sources of Information in a k-DPP Design

In the more general case, a k-DPP sampling design strategy could also allow for the inclusion of other potential extra sources of information. In the context of environmental monitoring networks, this could include topographic or demographic features. For instance, if our goal is to design a monitoring network for a given pollutant, we may consider demographic features correlated with the outcome of interest, such as population size, to diversify the location of the monitors. By using a k-DPP strategy, we would be more likely to choose locations in highly populated areas as well as not so populated ones.

Assuming that there exists p standardized features available about a given location $\mathbf{s}_i$, such that $(f^{(1)}(\mathbf{s}_i), \ldots, f^{(p)}(\mathbf{s}_i))$ is associated to $\mathbf{s}_i$, then

$$L_{i,j} = \exp\left\{ -\sum_{l=1}^{p} \frac{||f^{(l)}(\mathbf{s}_i) - f^{(l)}(\mathbf{s}_j)||^2}{2\sigma^2} \right\}. \tag{8.39}$$

We find this idea somewhat similar to Schmidt et al. (2011), where their goal was to handle nonstationarity by including the effect of covariates in the covariance structure of spatial processes. In this work, our aim is to select a diverse set of sampled design points by also taking into consideration the information available at the candidate points.

Furthermore, one may use an empirical description of the data available and define L as a sample covariance matrix.

Approximation to the Entropy-based Design Objective using a k-DPP Sampling Design Strategy

Another remarkable characteristic of a k-DPP design is its strong similarity to the entropy-based design objective in the Gaussian case, as discussed in the following result.

Result 8.6.2. *A k-DPP sampling design characterized by an L-ensemble given by the predictive covariance structure of the potential new monitors given the existing monitors can be viewed as a randomized version of the entropy design in the Gaussian case.*

In Section 8.8, we illustrate how a k-DPP sampling design strategy can be used as an approximation for the entropy-based designs for monitoring temperature fields.

8.7 Comparing k-DPP and SF Sampling Designs

Example 1. Beilshmiedia pendula trees in Barro Island

Let us consider as an example, the locations of Beilshmiedia pendula trees and elevation for a subset of an original survey plot in Barro Island. The data set `bei` is available for download in the `spatstat R` package (Baddeley and Turner, 2005; Baddeley et al., 2015). Figure 8.1 illustrates a total of 357 potential trees to be selected from. The objective is to select a subset of 20 of them.

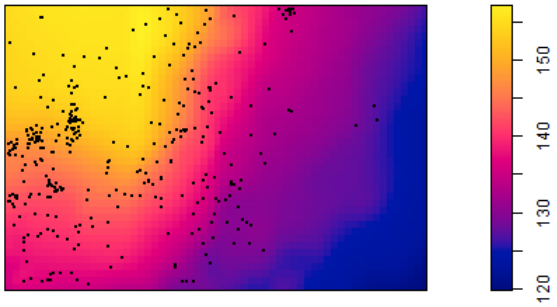


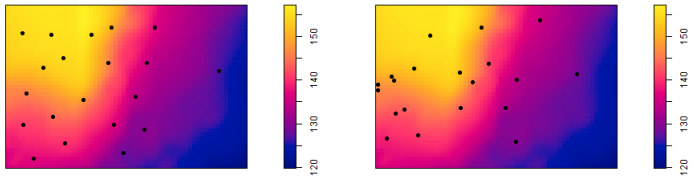
Figure 8.1: Tropical rainforest data. Locations of Beilshmiedia pendula trees and elevation (metres above sea level) in the $[700, 1000] \times [0, 200]$ metres window of a survey plot in Barro Island. Coloured background corresponds to the variation of elevation in that window, as seen in the scale on the right of the plot. Data available from `spatstat R` package.

Figure 8.2 illustrates the selected trees using the space-filling method

described in Section 8.3.1 as well as one based on a 20-DPP design. The 20-DPP design was characterized by an L-ensemble using a Gaussian kernel depended on both the locations and the elevation at all the potential trees, as in (8.39). Hence the k-DPP diversity was quantified by both the locations and the elevations. For illustrative purposes, we have assumed an arbitrary variance $\sigma^2 = 0.5$.

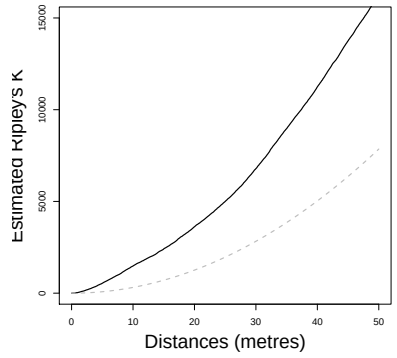
The space-filling design yielded a more spatially balanced design than the 20-DPP, as indicated in Figure 8.3. Nevertheless, the DPP design is not as spatially clustered as the candidate locations.

Note that even though our interest might be in a spatially-balanced design in order to better represent the study region, the DPP design will also "penalize" distant stations that are too much alike, hence also imposing diversity with respect to elevation. Here, we use the "penalization" term to reflect reduced probability of selection.

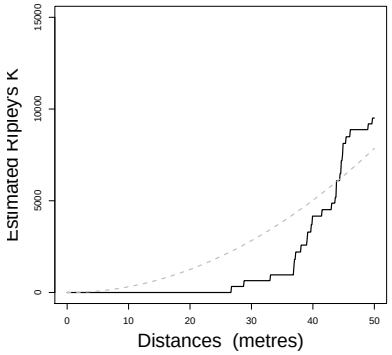


(a) Space-Filling (b) 20-DPP

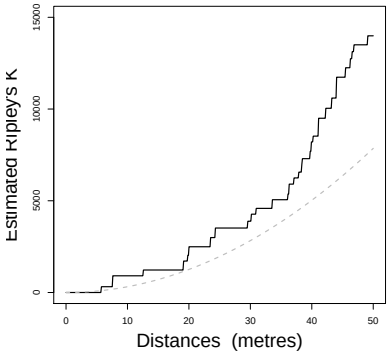
Figure 8.2: Locations of 20 Beilshmiedia pendula trees selected via a space-filling and a 20-DPP design strategies. Coloured background corresponds to the variation of elevation (metres above sea level).



(a) All candidates



(b) Space-Filling



(c) 20-DPP

Figure 8.3: Empirical (solid line) and theoretical (dashed line) Ripley's K. Note that the SF design shows more spatial regularity than the DPP design.

Example 2. Estimation Assessment with Artificial Data

In this study, we simulate a realization of a Matérn random field with mean zero, partial sill $\sigma^2 = 4$, range $\phi = 1$ and smoothness $\nu = 2$, as illustrated in Figure 8.4.

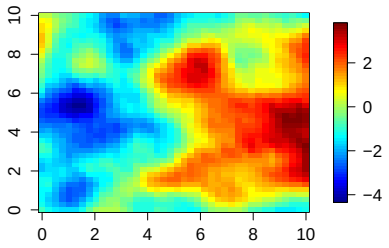


Figure 8.4: Realization of a Matérn random field with mean zero, partial sill $\sigma^2 = 4$, range $\phi = 1$ and smoothness $\nu = 2$, in a $[0, 10] \times [0, 10]$ domain. Coloured background corresponds to the observed values of the field (no units associated to them), as seen in the scale on the right of the plot.

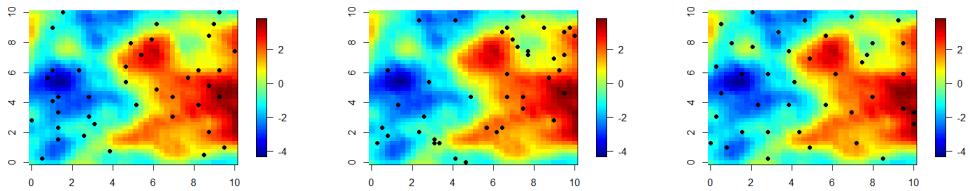
For the purposes of this simulation study, the realization was assumed to be an artificial true underlying field. We then repeatedly selected 40 locations using three design strategies: 40-DPP, via random uniform selection, and via a space-filling design. The 40-DPP was characterized using a Gaussian kernel with fixed variability of 0.5. No other source of information but the locations of the potential sites was assumed for any of the strategies considered. The potential sites consisted of a 40×40 fine grid over the spatial domain. Here, the objective is to compare the two spatially-balanced design strategies.

Figure 8.5 illustrates an example of the sampling locations using the different design strategies. Note that the space-filling design is not a randomized design strategy, so the variation in sampling locations is due to variation in variability in the start configuration points for the point swap algorithm. Note that these sampling locations would then emulate the environmental monitors and the observed value of the Matérn random field at those locations would reflect the data available.

In this study, we assumed no measurement error. The basic geostatistical model assumed that for each location $\mathbf{s}_i$ was

$$Y(\mathbf{s}_i) = \mu + \eta(\mathbf{s}_i), \tag{8.40}$$

$i = 1, \ldots, 40$, $\eta(\mathbf{s}_i)$ is a second-order stationary process with zero mean and partial sill σ^2, and Matérn correlation function with fixed smoothness $\nu = 2$.



(a) 40-DPP (b) Random (c) Space-Filling

Figure 8.5: Example of sampling locations using a 40-DPP design, random (uniform) selection, and a space-filling design.

We then proceeded with estimating the model parameters using the INLA method, as described in Section 3.2. To complete model specification, we assumed the following independent prior distributions for $\mathbf{\Psi} = (\mu, \sigma^2, \phi)$:

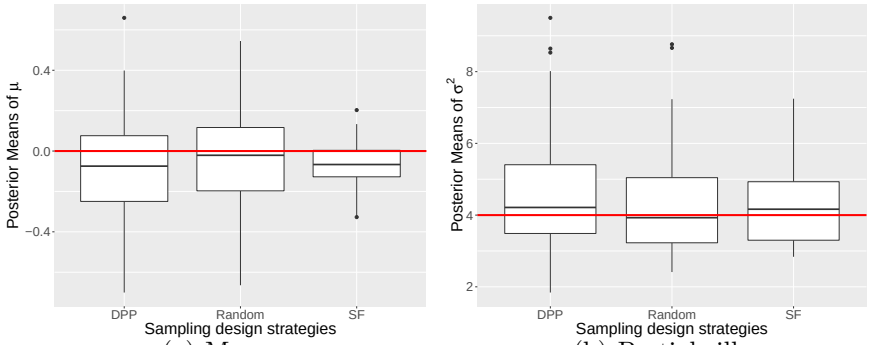
$$\mu \sim \mathcal{N}(0, 100) \tag{8.41}$$

$$\sigma^{-2} \sim \text{Gamma}(1, 0.01) \tag{8.42}$$

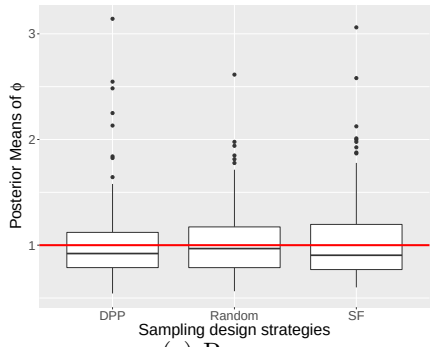
$$\phi \sim \text{Gamma}(1, 0.01). \tag{8.43}$$

These are fairly vague prior distributions. The notation above $\mathcal{N}$ denotes a normal with given mean and variance parameters.

Figure 8.6 contains the box-plots of the posterior means for these parameters, with posterior variability described in Figure 8.7. We have observed slightly more uncertainty *a posteriori* for the mean and partial sill parameters. On the other hand, it can be seen that for the vast majority of the simulations, the SF underestimated the true mean. The estimation performance of the DPP and SF design about the partial sill and range parameters seem comparable. When the main interest is in spatial regularity, in terms of estimation, we observe that the SF and 40-DPP design yield comparable design strategies.

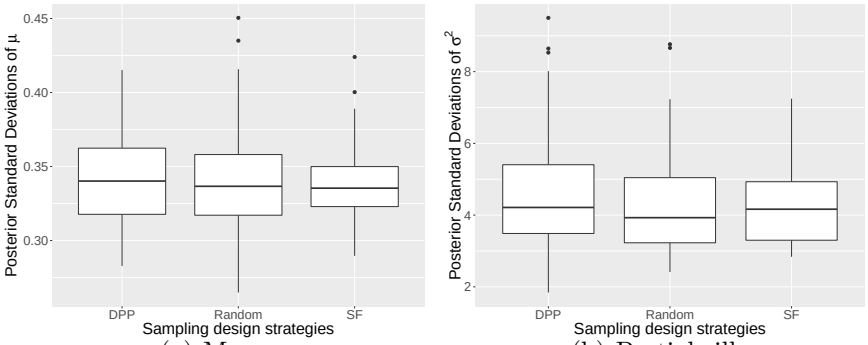


(a) Mean (b) Partial sill

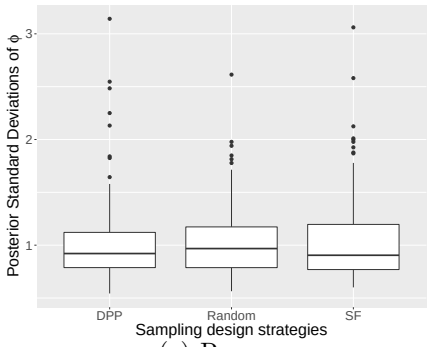


(c) Range

Figure 8.6: Box-plots of posterior means for the model parameters $\boldsymbol{\Psi} = (\mu, \sigma^2, \phi)$ after repeatedly selecting 40 locations using three different strategies: a 40-DPP with a Gaussian kernel, random uniform selection, and a space-filling design. This process was repeated 100 times. The red horizontal lines represent the true values of the parameters.



(a) Mean (b) Partial sill



(c) Range

Figure 8.7: Box-plots of posterior standard deviations for the model parameters $\boldsymbol{\Psi} = (\mu, \sigma^2, \phi)$ after repeatedly selecting 40 locations using three different strategies: a 40-DPP with a Gaussian kernel, random uniform selection, and a space-filling design. This process was repeated 100 times.

8.8 Comparing k-DPP and Entropy-Based Designs for Monitoring Temperature Fields

As noted in Section 4.1, temperature is now seen as an environmental hazard due to its potential negative effects in human health and nature. That leads to a need to ensure that the temperature field is adequately monitored. In this section, we compare through a brief case study, the designs obtained by the entropy and DPP approaches.

For this study we turn to the data described in Section 4.3.2, consisting of 97 stations spread over the Pacific Northwest where measurements of maximum daily temperature were obtained for the January to June 2000 period. A subset of 64 stations are to be selected among the 97 stations to constitute as a hypothetical monitoring network. An additional 33 stations are designated as potential sites for new monitors. In this case study, the goal is to select a subset of 10 stations from among the 33 to augment the network based on some design criterion.

From Section 4.4, recall that we have partitioned the data as $\mathbf{Y}_t \equiv (\mathbf{Y}_t^{(u)}, \mathbf{Y}_t^{(g)})$, $t = 1, \ldots, n$. Similarly, we can partition the data for a future time f as $\mathbf{Y}_f \equiv (\mathbf{Y}_f^{(u)}, \mathbf{Y}_f^{(g)})$. Denoting $\mathbf{y}_{1:n}^{(g)}$ as all the data available at the gauged sites across all times $1, \ldots, n$, then note that (Le and Zidek, 2006)

$$\mathbf{Y}_f^{(u)} | \mathbf{y}_f^{(g)}, \mathbf{y}_{1:n}^{(g)}, \mathbf{Z}, \mathbf{B}_0 \sim t_u \left(\boldsymbol{\mu}^{(u)}, \frac{d}{\delta - u + 1} \boldsymbol{\Xi}_{u|g}, \delta - u + 1 \right), \tag{8.44}$$

where

$$\begin{aligned} \boldsymbol{\mu}^{(u)} &= \mathbf{z}_f \mathbf{B}_0^{(u)} + \boldsymbol{\Xi}_{ug} \boldsymbol{\Xi}_{gg}^{-1} (\mathbf{y}_f^{(g)} - \mathbf{z}_f \mathbf{B}_0^{(g)}) & (8.45) \\ d &= 1 + \mathbf{z}_f \mathbf{F}^{-1} \mathbf{z}_f^\top + (\mathbf{y}_f^{(g)} - \mathbf{z}_f \mathbf{B}_0^{(g)}) \boldsymbol{\Xi}_{gg}^{-1} (\mathbf{y}_f^{(g)} - \mathbf{z}_f \mathbf{B}_0^{(g)})^\top & (8.46) \\ \boldsymbol{\Xi}_{u|g} &= \boldsymbol{\Xi}_{uu} - \boldsymbol{\Xi}_{ug} \boldsymbol{\Xi}_{gg}^{-1} \boldsymbol{\Xi}_{gu}. & (8.47) \end{aligned}$$

and that

$$\mathbf{Y}_f^{(g)} | \mathbf{y}_{1:t}^{(g)}, \mathbf{Z}, \mathbf{B}_0 \sim t_u \left(\boldsymbol{\mu}^{(g)}, \frac{c}{\delta + n - u - g + 1} \hat{\boldsymbol{\Xi}}_{gg}, l \right), \tag{8.48}$$

where

$$\begin{aligned}
\boldsymbol{\mu}^{(g)} &= (\mathbf{I}-\mathbf{W})\hat{\mathbf{B}}^{(g)} + \mathbf{W}\mathbf{B}_0^{(g)} && (8.49)\\
c &= 1 + \mathbf{z}_f(\mathbf{A}+\mathbf{F})^{-1}\mathbf{z}_f^\top && (8.50)\\
\mathbf{A} &= \sum_{t=1}^{n}\mathbf{z}_t^\top\mathbf{z}_t && (8.51)\\
\mathbf{W} &= (\mathbf{A}+\mathbf{F})^{-1}\mathbf{F}^{-1} && (8.52)\\
\hat{\boldsymbol{\Xi}}_{gg} &= \boldsymbol{\Xi}_{gg} + \mathbf{S} + (\hat{\mathbf{B}}^{(g)}-\mathbf{B}_0^{(g)})^\top(\mathbf{A}^{-1}+\mathbf{F}^{-1})^{-1}(\hat{\mathbf{B}}^{(g)}-\mathbf{B}_0^{(g)}) && (8.53)\\
\mathbf{S} &= \sum_{t=1}^{n}(\mathbf{y}_t^{(g)}-\mathbf{z}_t\hat{\mathbf{B}}^{(g)})^\top(\mathbf{y}_t^{(g)}-\mathbf{z}_t\hat{\mathbf{B}}^{(g)}) && (8.54)\\
\hat{\mathbf{B}}^{(g)} &= \left(\sum_{t=1}^{n}\mathbf{z}_t^\top\mathbf{z}_t\right)^{-1}\left(\sum_{t=1}^{n}\mathbf{z}_t^\top\mathbf{y}_t^{(g)}\right). && (8.55)
\end{aligned}$$

Using the results for the entropy of a multivariate t distribution described in Section 8.3.2, and denoting the available data as $\mathcal{D} = (\mathbf{y}_{1:n}^{(g)}, \mathbf{Z})$, conditionally on the hyperparameters, the total entropy can thus be decomposed as

$$H(\mathbf{Y}_f|\mathcal{D}) = H(\mathbf{Y}_f^{(u)}|\mathbf{Y}_f^{(g)},\mathcal{D}) + H(\mathbf{Y}_f^{(g)}|\mathcal{D}) \quad (8.56)$$

Proceeding as in the end of Section 8.3.2, recall that in the context of monitoring networks, when the goal is to augment the network, the objective is to find a subset of u^+ sites among the u ungauged ones (also referred to as candidate sites, where $\mathcal{C}$ denote the set of candidate points) to add to the existing network. We denote the remaining sites that are not the selected as u^-. The resulting network will then consist of $(\mathbf{Y}_f^{(u^+)}, \mathbf{Y}_f^{(g)})$.

Note that $\mathbf{Y}_f^{(u)}$ can be partitioned into $(\mathbf{Y}_f^{(u^+)}, \mathbf{Y}_f^{(u^-)})$. Thus,

$$\begin{aligned}
& H(\mathbf{Y}_f^{(u^+)}, \mathbf{Y}_f^{(u^-)}, \mathbf{Y}_f^{(g)}|\mathcal{D})\\
= \; & H(\mathbf{Y}_f^{(u^-)}|\mathbf{Y}_f^{(u^+)}, \mathbf{Y}_f^{(g)}, \mathcal{D}) + H(\mathbf{Y}_f^{(u^+)}, \mathbf{Y}_f^{(g)}|\mathcal{D}) && (8.57)
\end{aligned}$$

Notice that it will be optimal to augment the network with the u^+ sites so

as to maximize $H(\mathbf{Y}^{(u^+)}, \mathbf{Y}_f^{(g)}|\mathcal{D})$. Considering that

$$H(\mathbf{Y}^{(u^+)}, \mathbf{Y}_f^{(g)}|\mathcal{D}) = H(\mathbf{Y}_f^{(u^+)}|\mathbf{Y}_f^{(g)}, \mathcal{D}) + H(\mathbf{Y}_f^{(g)}|\mathcal{D}) \quad (8.58)$$

$$\overset{c}{\propto} \frac{1}{2}\log|\boldsymbol{\Xi}_{u^+|g}| + \frac{1}{2}\log\det(\hat{\boldsymbol{\Xi}}_{gg}), \quad (8.59)$$

then an equivalent criterion is to maximize $\frac{1}{2}\log\det(\boldsymbol{\Xi}_{u^+|g})$.

In summary, the entropy criterion for augmenting the network is thus

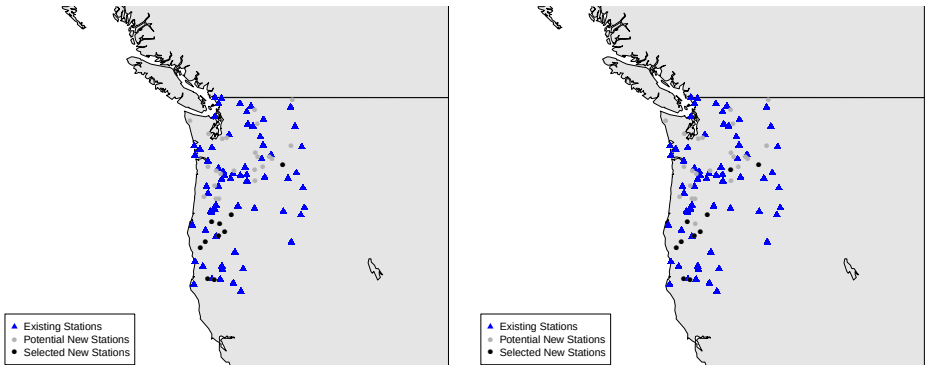
$$\arg\max_{u^+ \subset \mathcal{C}} \frac{1}{2}\log\det(\boldsymbol{\Xi}_{u|g}). \quad (8.60)$$

For the purposes of this case study, we considered an alternate 10-DPP design strategy characterized by the same hypercovariance matrix, $\boldsymbol{\Xi}_{u|g}$, in order to yield comparable design objectives. Similarly as in Section 4.4, we estimate it via the SG warping method based on the hypercovariance matrix among the gauged sites.

We obtained a 10-DPP design based on a simulation strategy of repeatedly sampling from a 10-DPP. Figure 8.8 illustrates the locations for selected stations for the two different methods. The entropy solution yielded a log-determinant of 78.70 for the optimal set of locations. Note that the majority of the new locations were selected in southern Oregon and northern California, west of the Cascade mountains. The k-DPP also selected a couple of sites in Eastern Washington, east of the Cascades.

Figure 8.10 illustrates the distribution of the log-determinants for the DPP samples considering different numbers of simulations. Moreover, from Figure 8.9, note that 10-DPP is very close to but suboptimal compared with the entropy design.

Though yielding a suboptimal solution, when the number of combinations is prohibitive, the simulation results indicate that sampling from a k-DPP could be used to obtain approximations to the entropy design. Further assessment to verify the properties of this method of approximating the entropy based design are needed, as the sampling complexity of the DPP will also increase for a prohibitive number of combinations.



(a) Entropy (b) 10-DPP

Figure 8.8: Comparison of entropy-based and DPP design strategies. The entropy solution yielded a log-determinant of 78.70 for the restricted conditional hypercovariance matrix for the ungauged sites considering the optimal set of locations. Here, we illustrate the solution of a 10-DPP sampling design strategy. Note the similarity in the choice of new locations across both designs.

Ko et al. (1995) use a branch-and-bound algorithm for this optimization problem, using a greedy strategy to obtain candidate sets of points. However, the algorithm is impractical for a large number of candidate points. On the other hand, Li et al. (2015) suggests an approximate sampling strategy for discrete k-DPPs that could be useful to alleviate the sampling complexity for a prohibitivele large number of combinations.

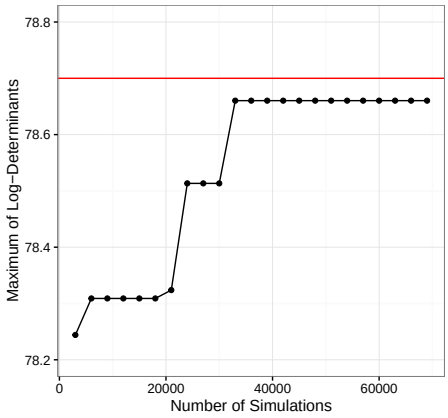
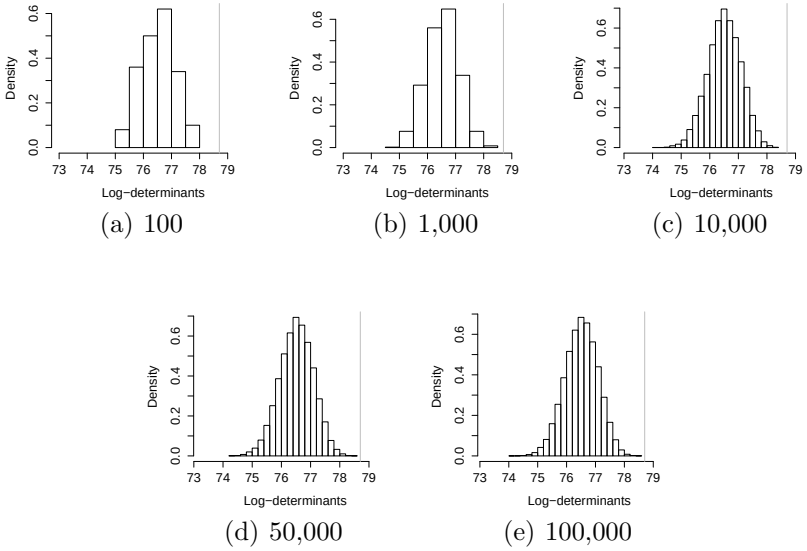


Figure 8.9: Current maximum log-determinants of the restricted conditional hypercovariance matrix for the ungauged sites when increasing the number of simulations of the 10-DPP.



(a) 100 (b) 1,000 (c) 10,000

(d) 50,000 (e) 100,000

Figure 8.10: Log-determinants of the restricted conditional hypercovariance matrix for the ungauged sites varying the number of simulations of a 10-DPP. The gray line represents the log-determinant for the optimal entropy solution.

8.9 Discussion and Future Work

We introduced a novel sampling design strategy based on k-determinantal point processes. The k-DPP design is a flexible design strategy that is able to yield spatially-balanced designs, while imposing additional diversity in the selection of locations based on additional features that might be available. The methodology is able to handle designing and redesigning a monitoring network. A summary of the important points discussed is as follows:

- The k-DPP design can be used as a spatially-balanced sampling design alternative to the space-filling design.
- The k-DPP design objective can be constructed in such a way that is strongly similar to the entropy-based design objective. It can thus be viewed as a randomized version of an entropy design.
- A sampling design strategy based on a k-DPP characterized by the same hypercovariance matrix of a entropy-based design optimal criterion, i.e. $\boldsymbol{\Xi}_{u|g}$, can be used as an approximation for the entropy-based design when the number of combinations is prohibitively large. Though suboptimal, this alternative could be particularly useful due to the NP-hardness of the entropy-based design optimal criterion.

Another randomized spatially balanced design is the generalized random tessellation stratified (GRTS) design (Stevens Jr and Olsen, 1999, 2003, 2004). For future work, we would like to investigate how the k-DPP design compares to the GRTS.

Moreover, in our studies we did not address inference about k-DPP L-ensemble parameters. Recent advances in a Bayesian framework include Affandi et al. (2014), which can serve as a starting point for future exploration. Our next step would be to investigate their methodology for learning parameters, and how these ideas could be used in the redesigning a monitoring network.

Chapter 9

Concluding Remarks

This thesis was motivated by the growing need for understanding the changes in Earth's climate as well as a increasing concerns due to their potential impact in human health. The focus was mostly on different aspects of temperature, particularly in the Pacific Northwestern region. Rapid changes and localized weather are very common in this region and the terrain plays an important role in separating often radically different climate and weather regimes.

When the goal is to understand a region's weather, we advocate performing exploratory analysis to better understand the localized changes in trend, instead of just focusing on the modelling of the spatial covariance structure. We argue that this is needed to better represent interesting smaller-scale trends, especially for regions with a complex terrain like the Pacific Northwest, which may not be captured by global climate models. We also extended the spatio-temporal model proposed in Le and Zidek (1992) and described how one could accommodate features in the mean that vary over space.

In addition, we explored the data fusion problem in order to combine information from multiple sources that might have been measured at different spatial scales. We saw that for environmental studies, data measurements are often supplemented by information brought by computer model outputs. We then introduced a scalable inference methodology using the INLA-SPDE approach, and illustrated this methodology for combining an ensemble of computer models output with measurements of temperature across the Pacific Northwest.

Another critical topic tackled was in designing monitoring networks. They play an important role in the surveillance of environmental processes. We introduced a novel flexible monitoring network design strategy based on

k-DPPs, and described how the methodology is able to handle both designing and redesigning a monitoring network. We illustrated how the k-DPP design is able to yield spatially-balanced designs, and could be used as a randomized design alternative to space-filling designs. Moreover, we noted that the k-DPP design objective is strongly similar to the entropy-based design one, and can be viewed as a randomized version of this design. Finally, we discussed how a sampling strategy based on k-DPPs could be particularly useful to approximate entropy-based design optimal solution when the number of combinations is prohibitive.

9.1 Future Work

In this section, we address the limitations of the methodologies presented in this thesis, and introduce potential alternatives, which are currently subject of future research.

9.1.1 Nonstationarity in `INLA-SPDE`: Inference for the BEM

In Chapter 5, we assumed a stationaty Matérn covariance structure for the "true" underlying random field. This assumption is somewhat unrealistic due to the complex terrain of the Pacific Northwest and its localized and abrupt changes in climate. We would therefore like to accommodate nonstationarity in the `INLA-SPDE` inference strategy. This would entail representing a Gaussian random field η as a solution to a SPDE with covariance parameters varying over space, and writing it as

$$(\kappa^2(\mathbf{s}) - \Delta)^{\frac{\alpha}{2}}\{\tau(\mathbf{s})\eta(\mathbf{s})\} = \mathcal{W}(\mathbf{s}), \tag{9.1}$$

where τ models the variance of the process and is allowed to vary on space.

We then aim to investigate whether this leads to improvement in the DBEM application of combining temperature measurements and an ensemble of deterministic model outputs for stations spread over the Pacific Northwest, as introduced Chapter 6.

9.1.2 Modified DBEM for Forecasting

In Chapter 6, we noted that the DBEM is based on a mixture of posterior distributions based on a training set. The methodology thus requires the computation of mixing weights, which are essentially based on normalized marginal likelihoods across the training set. Difficulty in differentiating these weights may be encountered when at least one log-likelihood is significantly higher than the rest, and ultimately will dominate the mixing weights. We introduced an alternative methodology described in Algorithm 2 that is the subject of current research.

9.1.3 Comparison of k-DPP Design with the Generalized Random Tessellation Stratified (GRTS) Design

In Chapter 8, we described how the k-DPP design can be used as a spatially balanced sampling design alternative to the space-filling design. Another randomized spatially balanced design is the generalized random tessellation stratified (GRTS) design (Stevens Jr and Olsen, 1999, 2003, 2004). How well the k-DPP design compares with the GRTS is the subject of future research.

9.1.4 Inference about k-DPP Design Parameters

In Chapter 8, we did not address inference about k-DPP L-ensemble parameters. Inference for these parameters could ultimately be used when redesigning a monitoring network, thus avoiding the need of an arbitrary selection. A starting point for exploration could be the work of Affandi et al. (2014). They proposed using MCMC methods, while focusing on random-walk Metropolis-Hastings (Metropolis et al., 1953; Hastings, 1970) and a slice sampler (Neal, 2003). Our next step would be to investigate their methodology for learning parameters, and how these ideas could be used in the redesigning a monitoring network, ultimately avoiding the need for an arbitrary selection.

Bibliography

Affandi, R. H., Fox, E. B., Adams, R. P. and B., T. (2014) Learning the Parameters of Determinantal Point Process Kernels. In *ICML*.

Affandi, R. H., Kulesza, A., Fox, E. and Taskar, B. (2013) Nyström Approximation for Large-Scale Determinantal Processes. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics*.

Affandi, R. H., Kulesza, A. and Fox, E. B. (2012) Markov determinantal point processes. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*.

Åström, C., Orru, H., Rocklöv, J., Strandberg, G., Ebi, K. L. and Forsberg, B. (2013) Heat-related respiratory hospital admissions in Europe in a changing climate: a health impact assessment. *BMJ Open*, **3**.

Baddeley, A., Rubak, E. and Turner, R. (2015) *Spatial Point Patterns: Methodology and Applications with R*. London: Chapman and Hall/CRC Press.

Baddeley, A. and Turner, R. (2005) spatstat: An R Package for Analyzing Spatial Point Patterns. *Journal of Statistical Software*, **12**, 1–42.

Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2014) *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC.

Bayarri, M. J. and DeGroot, M. H. (1989) Optimal Reporting of Predictions. *Journal of the American Statistical Association*, **84**, 214–222.

Berman, M. and Diggle, P. (1989) Estimating Weighted Integrals of the

Second-Order Intensity of a Spatial Point Process. *Journal of the Royal Statistical Society. Series B (Methodological)*, **51**, 81–92.

Berrocal, B., Gelfand, A. E. and Holland, D. M. (2010a) A Spatio-Temporal Downscaler for Output From Numerical Models. *Journal of Agricultural, Biological, and Environmental Statistics*, **15**, 176–197.

Berrocal, V. J., Gelfand, A. E. and Holland, D. M. (2010b) A bivariate variate space-time downscaler under space and time misaligment. *The Annals of Applied Statistics*, **4**, 1942–1975.

— (2012) Space-Time Data fusion Under Error in Computer Model Output: An Application to Modeling Air Quality. *Biometrics*, **68**, 837–848.

Berrocal, V. J., Raftery, A. E. and Gneiting, T. (2007) Combining Spatial Statistical and Ensemble Information in Probabilistic Weather Forecasts. *Monthly Weather Review*, **135**, 1386–1402.

Bivand, R. S., Pebesma, E. and Gómez-Rubio, V. (2013) *Applied Spatial Data Analysis with R*, chap. Spatial Point Pattern Analysis, 173–211. New York, NY: Springer New York.

Blangiardo, M. and Cameletti, M. (2015) *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons.

Bornn, L., Shaddick, G. and Zidek, J. V. (2012) Modeling Nonstationary Processes Through Dimension Expansion. *Journal of the American Statistical Association*, **107**, 281–289.

Borodin, A. (2009) Determinantal Point Processes.

Borodin, A. and Olshanski, G. (2000) Distributions on Partitions, Point Processes, and the Hypergeometric Kernel. *Communications in Mathematical Physics*, **211**, 335–358.

Borodin, A. and Rains, E. M. (2005) Eynard-Mehta Theorem, Schur Process, and their Pfaffian Analogs. *Journal of Statistical Physics*, **121**, 291–317.

Cameletti, M., Lindgren, F., Simpson, D. and Rue, H. (2013) Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *AStA Advances in Statistical Analysis*, **97**, 109–131.

Caselton, W. F., Kan, L. and Zidek, J. V. (1992) *Statistics in the Environmental & Earth Sciences*, chap. Quality data networks that minimize entropy. A Hodder Arnold Publication. E. Arnold.

Cox, D. D., Cox, L. H. and Ensor, K. B. (1997) Spatial sampling and the environment: some issues and directions. *Environmental and Ecological Statistics*, **4**, 219–233.

Cressie, N. and Wikle, C. (2011) *Statistics for Spatial Data*. J. Wiley.

Cressie, N. A. C. (1993) *Statistics for Spatial Data*. J. Wiley.

Daly, C., Halbleib, M., Smith, J. I., Gibson, W. P., Doggett, M. K., Taylor, G. H., Curtis, J. and Pasteris, P. P. (2008) Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *International Journal of Climatology*, **28**, 2031–2064.

Daly, C., Neilson, R. P. and Phillips, D. L. (1994) A Statistical-Topographic Model for Mapping Climatological Precipitation over Mountainous Terrain. *Journal of Applied Meteorology*, **33**.

Daly, C., Taylor, G. and Gibson, W. (1997) The PRISM Approach to Mapping Precipitation and Temperature. In *10th Conf. on Applied Climatology, Reno, NV, Amer. Meteor. Soc.*, no. 10-12.

Daly, C., Taylor, G., Gibson, W., Parzybok, T., Johnson, G. and Pasteris, P. (2000) High-quality spatial climate data sets for the United States and beyond. *Transactions of the ASAE*, **43**, 1957–1962.

Damian, D., Sampson, P. D. and Guttorp, P. (2001) Bayesian estimation of semi-parametric non-stationary spatial covariance structures. *Environmetrics*, **12**, 161–178.

De Oliveira, V., Kedem, B. and Short, D. A. (1997) Bayesian Prediction of Transformed Gaussian Random Fields. *Journal of the American Statistical Association*, **92**, 1422–1433.

Diggle, P. (1985) A Kernel Method for Smoothing Point Process Data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **34**, 138–147.

Diggle, P. J. . (2013) *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*. Chapman and Hall/CRC, 3rd edn.

Diggle, P. J., Menezes, R. and Su, T.-l. (2010) Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **59**, 191–232.

Diggle, P. J. and Ribeiro Jr, P. J. (2007) *Model-based Geostatistics (Springer Series in Statistics)*. Springer, 1 edn.

Diggle, P. J., Tawn, J. A. and Moyeed, R. A. (1998) Model-based Geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **47**, 299350.

Ferreira, G. S. and Gamerman, D. (2015) Optimal Design in Geostatistics under Preferential Sampling. *Bayesian Anal.*, **10**, 711–735.

Foley, K. M. and Fuentes, M. (2008) A statistical framework to combine multivariate spatial data and physical models for Hurricane surface wind prediction. *Journal of Agricultural, Biological, and Environmental Statistics*, **13**, 37–59.

Fraley, C., Raftery, A. E., Gneiting, T. and M., S. J. (2013) ensembleBMA: An R Package for Probabilistic Forecasting using Ensembles and Bayesian Model Averaging. *Tech. Rep. 516*, University of Washington.

Fuentes, M. (2001) A high frequency kriging approach for non-stationary environmental processes. *Environmetrics*, **12**, 469–483.

— (2002) Spectral methods for nonstationary spatial processes. *Biometrika*, **89**, 197–210.

Fuentes, M. and Raftery, A. E. (2005) Model Evaluation and Spatial Interpolation by Bayesian Combination of Observations with Outputs from Numerical Models. *Biometrics*, **61**, 36–45.

Fuentes, M. and Smith, R. L. (2001) A New Class of Nonstationary Spatial Models. *Tech. rep.*, North Carolina State University.

Gelfand, A. E. (2010) *Handbook of Spatial Statistics*, chap. Misaligned spatial data: The change of support problem, 517–539. CRC Press.

Gelfand, A. E., Sahu, S. K. and Holland, D. M. (2012) On the effect of preferential sampling in spatial prediction. *Environmetrics*, **23**, 565–578.

Gelfand, A. E. and Schliep, E. M. (2016) Spatial statistics and Gaussian processes: A beautiful marriage. *Spatial Statistics*.

Gelfand, A. E. and Smith, A. F. M. (1990) Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, **85**, 398–409.

Gillenwater, J., Kulesza, A. and Taskar, B. (2012) Near-Optimal MAP Inference for Determinantal Point Processes. In *Advances in Neural Information Processing Systems*.

Goodchild, M. F. and Haining, R. P. (2004) GIS and spatial data analysis: Converging perspectives. *Papers in Regional Science*, **83**, 363–385.

Gotway, C. A. and Young, L. J. (2002) Combining Incompatible Spatial Data. *Journal of the American Statistical Association*, **97**, 632–648.

Grimit, E. P. and Mass, C. F. (2002) Initial Results of a Mesoscale Short-Range Ensemble Forecasting System over the Pacific Northwest. *Weather and Forecasting*, **17**, 192–205.

Guillas, S., Bao, J., Choi, Y. and Wang, Y. (2008) Statistical correction and downscaling of chemical transport model ozone forecasts over Atlanta. *Atmospheric Environment*, **42**, 1338 – 1348.

Guillas, S., Tiao, G., Wuebbles, D. and Zubrow, A. (2006) Statistical diagnostic and correction of a chemistry-transport model for the prediction of total column ozone. *Atmospheric Chemistry and Physics*, **6**, 525–537.

Guttorp, P., Le, N. D., Sampson, P. D. and Zidek, J. V. (1993) Using entropy in the redesign of an environmental monitoring network. *Tech. rep.*, Technical report, Department of Statistics. University of British Columbia., 1992. Tech. Rep. 116.

Guttorp, P. and Sampson, P. D. (2010) Discussion of Geostatistical inference under preferential sampling by Diggle, P. J., Menezes, R. and Su, T. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **59**, 191–232.

Haaland, P., McMillan, N., Nychka, D. and Welch, W. (1994) Analysis of Space-Filling Designs. vol. 26, 111–120.

Haas, T. C. (1990) Lognormal and Moving Window Methods of Estimating Acid Deposition. *Journal of the American Statistical Association*, **85**, pp. 950–963.

— (1995) Local Prediction of a Spatio-Temporal Process with an Application to Wet Sulfate Deposition. *Journal of the American Statistical Association*, **90**, pp. 1189–1199.

Hastings, H. (1970) Monte Carlo Sampling Methods Using Markov chains and their Applications. *Biometrika*, **57**, 97–109.

Higdon, D., Swall, J. and Kern, J. (1999) Non-stationary spatial modeling. *Bayesian statistics*, **6**, 761–768.

Hoberg, E. P. and Brooks, D. R. (2015) Evolution in action: climate change, biodiversity dynamics and emerging infectious disease. *Philosoph-*

ical Transactions of the Royal Society of London B: Biological Sciences, **370**.

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999) Bayesian model averaging: a tutorial (with comments by M. Clyde, David Draper and E. I. George, and a rejoinder by the authors. *Statist. Sci.*, **14**, 382–417.

Hough, J. B., Krishnapur, M., Peres, Y. and Virg, B. (2006) Determinantal Processes and Independence. *Probab. Surveys*, **3**, 206–229.

Jaynes, E. T. (1963) *Blendeis University Summer Institute Lectures in Theoretical Physics, Statistical Physics 3*, chap. Information theory and statistical mechanics, 181–218. New York: W. A. Benjamin, Inc.

Kahle, D. and Wickham, H. (2013) ggmap: Spatial Visualization with ggplot2. *The R Journal*, **5**, 144–161.

Kennedy, M. C. and O'Hagan, A. (2001) Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **63**, 425–464.

Kleiber, W., Katz, R. W. and Rajagopalan, B. (2013) Daily minimum and maximum temperature simulation over complex terrain. *The Annals of Applied Statistics*, **7**, 588–612.

Kleiber, W., Raftery, A. E., Baars, J., Gneiting, T., Mass, C. F. and Grimit, E. (2011) Locally Calibrated Probabilistic Temperature Forecasting Using Geostatistical Model Averaging and Local Bayesian Model Averaging. *Monthly Weather Review*, **139**, 2630–2649.

Ko, C.-W., Lee, J. and Queyranne, M. (1995) An Exact Algorithm for Maximum Entropy Sampling. *Operations Research*, **43**, pp. 684–691.

Kulesza, A. and Taskar, B. (2011a) Learning Determinantal Point Processes. In *Conference on Uncertainty in Artificial Intelligence (UAI)*. Barcelona, Spain.

— (2011b) Structured Determinantal Point Processes. In *Advances in Neural Information Processing Systems 23*.

— (2012) Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, **5**.

Kunkel, K. E., Stevens, L. E., Stevens, S. E., Sun, L., Janssen, E., Wuebbles, D., Redmond, K. T. and Dobson, J. G. (2013) Regional Climate Trends and Scenarios for the U.S. National Climate Assessment. Part 6. Climate of the Northwest U.S. *Tech. rep.*, U.S. NOAA Technical Report NESDIS 142-6. National Oceanic and Atmospheric Administration, National Environmental Satellite, Data, and Information Service, Washington, D.C.

Lavancier, F., Møller, J. and Rubak, E. (2015) Determinantal point process models and statistical inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Lawrimore, J. H., Menne, M. J., Gleason, B. E., Williams, C. N., Wuertz, D. B., Vose, R. S. and Rennie, J. (2011) An overview of the Global Historical Climatology Network monthly mean temperature data set, version 3. *Journal of Geophysical Research: Atmospheres*, **116**.

Le, N., Zidek, J., White, R., Cubranic, D., with Fortran code for Sampson-Guttorp estimation authored by Paul D. Sampson, Guttorp, P., Meiring, W., Hurley, C., method implementation by H.A. Watts, R.-K.-F. and Shampine., L. (2014) *EnviroStat: Statistical analysis of environmental space-time processes*. R package version 0.4-0.

Le, N. D. and Zidek, J. (2006) *Statistical Analysis of Environmental Space-Time Processes (Springer Series in Statistics)*. Springer, 1 edn edn.

Le, N. D. and Zidek, J. V. (1992) Interpolation with uncertain spatial covariances: A Bayesian alternative to Kriging. *Journal of Multivariate Analysis*, **43**, 351 – 374.

Li, B., Sain, S., Mearns, L. O., Anderson, H. A., Kovats, S., Ebi, K. L., Bekkedal, M. Y. V., Kanarek, M. S. and Patz, J. A. (2012) The impact

of extreme heat on morbidity in Milwaukee, Wisconsin. *Climatic Change*, **110**, 959–976.

Li, C., Jegelka, S. and Sra, S. (2015) Efficient sampling for k-determinantal point processes. *arXiv preprint arXiv:1509.01618*.

Lindgren, F. and Rue, H. (2015) Bayesian Spatial Modelling with R-INLA. *Journal of Statistical Software*, **63**, 1–25.

Lindgren, F., Rue, H. and Lindström, J. (2011) An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**, 423–498.

Liu, Z. (2007) *Combining measurements with deterministic model outputs: predicting ground-level ozone*. Ph.D. thesis, University of British Columbia.

Liu, Z., Le, N. D. and Zidek, J. V. (2011) An empirical assessment of Bayesian melding for mapping ozone pollution. *Environmetrics*, **22**, 340–353.

Macchi, O. (1975) The Coincidence Approach to Stochastic Point Processes. *Advances in Applied Probability*, **7**, pp. 83–122.

Mass, C. (2008) *The Weather of the Pacific Northwest*. University of Washington Press, 1 edn.

McMillan, N. J., Holland, D. M., Morara, M. and Feng, J. (2010) Combining numerical model output and particulate data using Bayesian space–time modeling. *Environmetrics*, **21**, 48–65.

Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E. and Houston, T. G. (2012) An Overview of the Global Historical Climatology Network-Daily Database. *Journal of Atmospheric and Oceanic Technology*, **29**, 897910.

Metropolis, N., Rosenbulth, A., Rosenbulth, M., Teller, A. and Teller, E. (1953) Equation of State Calculations by Fast Computing Machine. *Journal of Chemical Physics*, **21**, 1087–1091.

Møller, J. and Waagepetersen, R. P. (2004) *Statistical inference and simulation for spatial point processes*. Chapman & Hall/CRC.

Mote, P. (2004) " How and why is Northwest climate changing?" in Climate Change, Carbon, and Forestry in Northwestern North America.

Mote, P., Snover, A. K., Capalbo, S., Eigenbrode, S. D., Glick, P., Littell, J., Raymondi, R. and Reeder, S. (2014) Ch. 21: Northwest. Climate Change Impacts in the United States: The Third National Climate Assessment. U.S. Global Change Research Program.

Müller, W. G. (2005) A comparison of spatial design methods for correlated observations. *Environmetrics*, **16**, 495–505.

Murray, A. T. (2010) Advances in location modeling: GIS linkages and contributions. *Journal of Geographical Systems*, **12**, 335–354.

Naylor, J. C. and Smith, A. F. M. (1982) Applications of a Method for the Efficient Computation of Posterior Distributions. *Journal of the Royal Statistical Society Series C*, **31**, 214–225.

Neal, R. M. (2003) Slice sampling. *Ann. Statist.*, **31**, 705–767.

Nguyen, H., Cressie, N. and Braverman, A. (2012) Spatial Statistical Data Fusion for Remote Sensing Applications. *Journal of the American Statistical Association*, **107**, 1004–1018.

Nychka, D., Furrer, R., Paige, J. and Sain, S. (2015) fields: Tools for spatial data. R package version 8.4-1.

Nychka, D. and Saltzman, N. (1998) *Case Studies for Environmental Statistics*, chap. Design of air quality networks. Lecture Notes in Statistics, Springer Verlag.

Nychka, D., Wikle, C. and Royle, J. A. (2002) Multiresolution models for nonstationary spatial covariance functions. *Statistical Modelling*, **2**, 315–331.

Nychka, D., Yang, Q. and Royle, J. A. (1997) *Statistics for the Environment, Vol. 3, Pollution Assessment and Control*, chap. Constructing spatial designs using regression subset selection. Wiley, New York.

Nychka, D. W. and Anderson, J. L. (2010) *Handbook of Spatial Statistics*, chap. Data assimilation, 241–270. CRC Press.

Palacios, M. B. and Steel, M. F. J. (2006) Non-gaussian bayesian geostatistical modelling. *Journal of American Statistical Association*, **101**, 604618.

Pati, D., Reich, B. J. and Dunson, D. B. (2011) Bayesian geostatistical modelling with informative sampling locations. *Biometrika*, **98**, 35–48.

Pronzato, L. and Müller, W. G. (2012) Design of computer experiments: space filling and beyond. *Statistics and Computing*, **22**, 681–701.

R Core Team (2014) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raftery, A. E., Balabdaoui, F., Gneiting, T. and Polakowski, M. (2005) Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, **133**, 1155–1174.

Raftery, A. E., Madigan, D. and Hoeting, J. A. (1997) Bayesian Model Averaging for Linear Regression Models. *Journal of the American Statistical Association*, **92**, 179–191.

Ribeiro Jr., P. and Diggle, P. (2001) geoR: a package for geostatistical analysis. *R-NEWS*, **1**, 15–18.

Ripley, B. D. (1988) *Statistical inference for spatial processes*. Cambridge University Press.

Robine, J.-M., Cheung, S. L. K., Roy, S. L., Oyen, H. V., Griffiths, C., Michel, J.-P. and Herrmann, F. R. (2008) Death toll exceeded 70,000 in Europe during the summer of 2003. *Comptes Rendus Biologies*, **331**, 171 – 178.

Rodriguez-Mozaz, S., Lopez de Alda, M. J. and Barceló, D. (2006) Biosensors as useful tools for environmental analysis and monitoring. *Analytical and Bioanalytical Chemistry*, **386**, 1025–1041.

Royle, J. A. and Nychka, D. (1998) An Algorithm for the Construction of Spatial Coverage Designs with Implementation in SPLUS. *Comput. Geosci.*, **24**, 479–488.

Rue, H. and Martino, S. (2007) Approximate Bayesian inference for hierarchical Gaussian Markov random field models. *Journal of statistical planning and inference*, **137(10)**, 3177–3192.

Rue, H., Martino, S. and Chopin, N. (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71**, 319–392.

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P. and Lindgren, F. K. (2016) Bayesian Computing with INLA: A Review. *arXiv preprint arXiv:1604.00860*.

Ruiz-Cárdenas, R., Krainski, E. T. and Rue, H. (2012) Direct fitting of dynamic models using integrated nested Laplace approximations {INLA}. *Computational Statistics & Data Analysis*, **56**, 1808 – 1828.

Sacks, J., Welch, W. J., Mitchell, T. J. and Wynn, H. P. (1989) Design and Analysis of Computer Experiments. *Statist. Sci.*, **4**, 409–423.

Salathé, E. P., Steed, R., Mass, C. F. and Zahn, P. H. (2008) A High-Resolution Climate Model for the U.S. Pacific Northwest: Mesoscale Feedbacks and Local Responses to Climate Change. *J. Climate*, **21**, 57085726.

Sampson, P. D. and Guttorp, P. (1992) Nonparametric Estimation of Nonstationary Spatial Covariance Structure. *Journal of the American Statistical Association*, **87**, 108–119.

Schmidt, A. M., Guttorp, P. and O'Hagan, A. (2011) Considering covariates in the covariance structure of spatial processes. *Environmetrics*, **22**, 487–500.

Schmidt, A. M. and O'Hagan, A. (2003) Bayesian inference for nonstationary spatial covariance structure via spatial deformations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **65**, 743–758.

Shirota, S. and Gelfand, A. E. (2016) Approximate Bayesian Computation and Model Validation for Repulsive Spatial Point Processes. *ArXiv e-prints*.

Simpson, D., Lindgren, F. and Rue, H. (2012a) In order to make spatial statistics computationally feasible, we need to forget about the covariance function. *Environmetrics*, **23**, 65–74.

— (2012b) Think continuous: Markovian Gaussian models in spatial statistics. *Spatial Statistics*, **1**, 16 – 29.

Smith, A. F. M., Skene, A. M., Shaw, J. E. H. and Naylor, J. C. (1987) Progress with Numerical and Graphical Methods for Practical Bayesian Statistics. *Journal of the Royal Statistical Society. Series D (The Statistician)*, **36**, 75–82.

Smith, B. J. and Cowles, M. K. (2007) Correlating point-referenced radon and areal uranium data arising from a common spatial process. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **56**, 313–326.

Snyder, J. P. (1987) *Map projections–a working manual*. United States.

Stevens Jr, D. L. and Olsen, A. R. (2003) Variance estimation for spatially balanced samples of environmental resources. *Environmetrics*, **14**, 593–610.

— (2004) Spatially balanced sampling of natural resources. *Journal of the American Statistical Association*, **99**, 262–278.

Stevens Jr, D. L. and Olsen, A, R. (1999) Spatially restricted surveys over time for aquatic resources. *Journal of Agricultural, Biological, and Environmental Statistics*, 415–428.

Strauss, W. A. (2008) *Partial Differential equations – An Introduction*. New York: John Wiley and Sons, Inc.

Swall, J. L. and Davis, J. M. (2006) A Bayesian statistical approach for the evaluation of {CMAQ}. *Atmospheric Environment*, **40**, 4883 – 4893. Special issue on Model Evaluation: Evaluation of Urban and Regional Eulerian Air Quality Models.

Tierney, L. and Kadane, J. B. (1986) Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistics Association*, **81**, 82–86.

Tierney, L., Kass, R. E. and Kadane, J. B. (1989) Approximate Marginal Densities of Nonlinear Functions. *Biometrika*, **76**, 425–433.

Tothill, I. E. (2001) Biosensors developments and potential applications in the agricultural diagnosis sector. *Computers and Electronics in Agriculture*, **30**, 205 – 218.

Van Groenigen, J., Pieters, G. and Stein, A. (2000) Optimizing spatial sampling for multivariate contamination in urban areas. *Environmetrics*, **11**, 227–244.

Whittle, P. (1954) ON STATIONARY PROCESSES IN THE PLANE. *Biometrika*, **41**, 434–449.

— (1963) Stochastic-processes in several dimensions. *Bulletin of the International Statistical Institute*, **40**, 974–994.

Wikle, C. K. and Berliner, L. M. (2005) Combining Information across Spatial Scales. *Technometrics*, **47**, 80–91.

World Health Organization (2009) Global health risks: mortality and burden of disease attributable to selected major risks. URLhttp://www.who.int/healthinfo/global_burden_disease/GlobalHealthRisks_report_full.pdf. Last Accessed: 2016-06-06.

Zidek, J. V., Le, N. D. and Liu, Z. (2012) Combining data and simulated data for space-time fields: application to ozone. *Environmental and Ecological Statistics*, **19**, 37–56.

Zidek, J. V., Shaddick, G. and Taylor, C. G. (2014) Reducing estimation bias in adaptively changing monitoring networks with preferential site selection. *Ann. Appl. Stat.*, **8**, 1640–1670.

Zidek, J. V. and Zimmerman, D. L. (2010) *Handbook of Spatial Statistics*, chap. Monitoring Network Design, 131–148. CRC Press.

Appendix A

Miscellaneous

Definition A.1. ***Green's first identity*** *(Chapter 7 of Strauss (2008))* *Let u and v be scalar functions on some region $\Omega \subset \mathbb{R}^d$. Suppose that u is twice continuously differentiable, and v once continuously differentiable. Then,*

$$\int_{\partial\Omega} v\frac{\partial u}{\partial n}dS = \int_{\Omega} \nabla v \cdot \nabla u \; d\mathbf{x} + \int_{\Omega} v\Delta u \; d\mathbf{x}, \tag{A.1}$$

where Δ is the Laplacian operator, $\partial\Omega$ is the boundary region Ω, and $\frac{\partial u}{\partial n}$ is the directional derivative in the outward normal direction.

Appendix B

INLA-SPDE Example

In this appendix, we demonstrate the use of `R-INLA` for the implementation of the BEM model described in Chapter 5. We consider an artificial ensemble of five deterministic model outputs. We randomly sampled 100 locations within a unit square, and simulated data based on the following settings.

- **Parameters of the "true" underlying process Z**
  - **Mean parameters**:

$$\mu(\mathbf{s}) = \alpha + \beta_1 \text{lat} + \beta_2 \text{long},$$

    where $\alpha = 1$, $\beta_1 = 2$, $\beta_2 = 1$.
  - **Covariance parameters**: Matérn with

    smoothness $\kappa = 10$;
    marginal variance $\sigma^2 = 1$; and
    range $\sqrt{8}/\kappa = 0.28$

- **Parameters of measurement process $\hat{\mathbf{Z}}$**

$$\sigma_e^2 = 0.5.$$

- **Parameters of ensemble members $\tilde{\mathbf{Z}}_j$**
  - Additive biases $(a_1, a_2, a_3, a_4, a_5) = (2.0, 3.5, 1.5, 2.5, 3.0)$.
  - Multiplicative biases $(b_1, b_2, b_3, b_4, b_5) = (0.8, 1.2, 0.9, 1.1, 1.5)$.
  - Variances $(\sigma^2_{\delta_1}, \sigma^2_{\delta_2}, \sigma^2_{\delta_3}, \sigma^2_{\delta_4}, \sigma^2_{\delta_5}) = (2.0, 2.5, 1.5, 2.0, 3.0)$.

After simulating the data, the first step is to create a triangulation of the continuous spatial domain, as it is described in Figure B.1. Note that the spatial domain was extended to avoid a boundary effect.

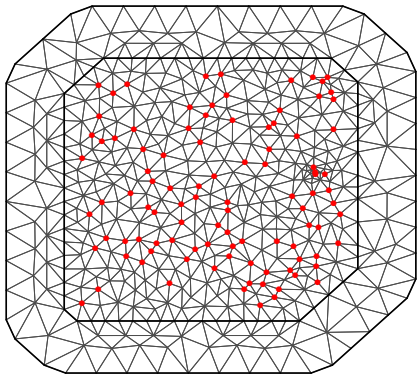


Figure B.1: Triangulation for the artificial data. The mesh comprises of 486 edges and was constructed using triangles that have a minimum angle of 25, maximum edge length of 0.1 within the spatial domain and 0.2 in the extension domain. The 100 artificial monitoring locations are highlighted in red.

For model completeness, in order to carry out the inference procedure, we describe our independent prior specifications below. For numerical stability, we specify priors for precision parameters (inverse of the variance) in a logarithmic scale. Notation used below for normal priors is mean and precision, whereas we refer to a Log-gamma as simply the logarithm of a Gamma distribution.

- For the mean parameters α, β_1, β_2, and calibration parameters a_j and b_j, for $j = 1, \ldots, 5$, we specified a $N(0, 0.01)$ prior.
- For $\log(\sigma_{\delta_j}^{-2})$, $j = 1, \ldots, 5$, we specified a Log-gamma$(0.01, 0.01)$ prior.
- For $\log(\sigma_e^{-2})$, we specified a Log-gamma$(1, 0.01)$ prior.

- For $\log(\sigma)$, we specify a $N(0, 0.1)$ prior, for $\log(\kappa)$ a $N(0, 1)$. We heuristically specify the prior for the spatial range as a fifth of the approximate domain diameter. This leads to a fairly vague prior specification for $\log(\sigma)$. As described in Lindgren and Rue (2015), for this heuristic choice, the precision 1 for the prior of $\log(\kappa)$ gives an approximate 95% prior probability for the range being shorter than the domain size.

Table B.1 contains the posterior summaries for the parameters of the BEM model for the artificial data, all parameters were reasonably well estimated. Under a fairly weak prior specification, the measurement error precision was underestimated. The marginal posterior densities for all model parameters can be found in Figures B.2, B.3, B.4, B.5, B.6, and B.7.

Table B.1: Posterior summaries for the parameters of the BEM model for the artificial data, including posterior mean and 95% credible intervals. Note that all parameters were reasonably well estimated. Under a fairly weak prior specification, the measurement error precision was underestimated.

Parameter	True Value	Post. Mean	CI (95%)
α	1.00	0.87	(-0.78; 2.56)
β_1	2.00	1.55	(-0.61; 3.60)
β_2	1.00	2.36	(0.29; 4.62)
σ_e^{-2}	2.00	3.38	(2.07; 5.27)
σ^{-2}	1.00	1.40	(0.65; 2.56)
κ	10.0	8.76	(4.63; 13.52)
ρ	0.28	0.32	(0.18; 0.53)
a_1	2.00	1.81	(1.15; 2.45)
a_2	3.50	3.25	(2.49; 4.01)
a_3	1.50	1.06	(0.44; 1.65)
a_4	2.50	2.55	(1.85; 3.26)
a_5	3.00	2.93	(2.03; 3.85)
b_1	0.80	0.91	(0.71; 1.11)
b_2	1.20	1.30	(1.07; 1.53)
b_3	0.90	1.04	(0.85; 1.24)
b_4	1.10	1.14	(0.92; 1.36)
b_5	1.50	1.45	(1.17; 1.73)
$\sigma_{\delta_1}^{-2}$	0.50	0.64	(0.47; 0.85)
$\sigma_{\delta_2}^{-2}$	0.40	0.47	(0.34; 0.63)
$\sigma_{\delta_3}^{-2}$	0.67	0.70	(0.51; 0.94)
$\sigma_{\delta_4}^{-2}$	0.50	0.53	(0.39; 0.72)
$\sigma_{\delta_5}^{-2}$	0.33	0.28	(0.21; 0.38)

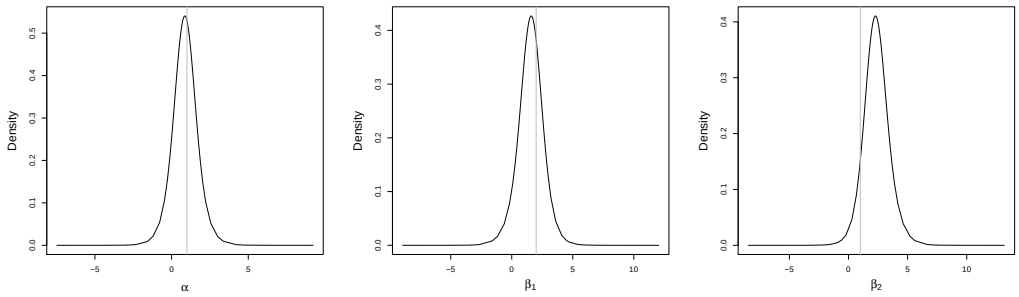


Figure B.2: Posterior distributions for the mean parameters of the underlying random field. The gray line represents the true value.

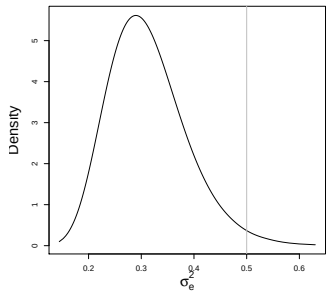


Figure B.3: Posterior distribution for the measurement error variance i.e, σ_e^2. The gray line represents the true value.

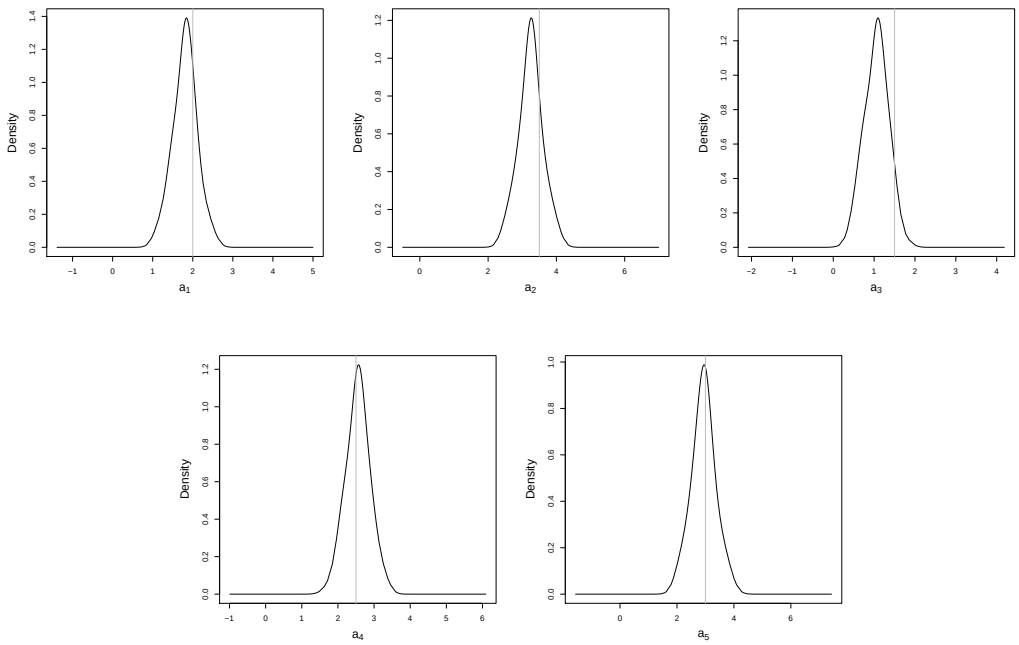


Figure B.4: Posterior distributions for the additive calibration parameters for each member of the ensemble i.e, a_j, for $j = 1, \ldots, 5$. The gray line represents the true value.

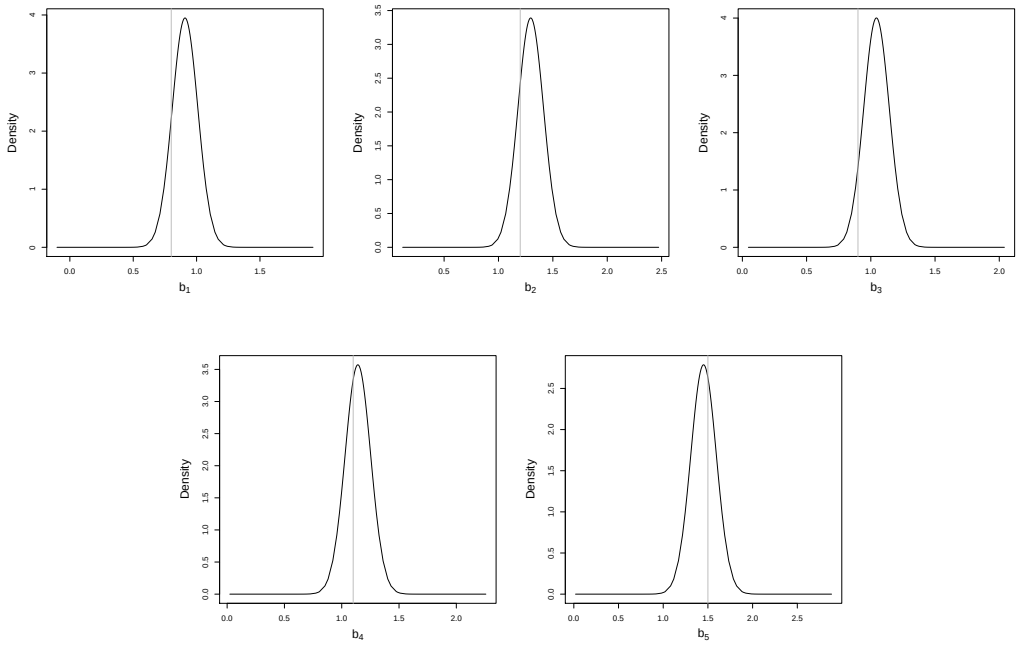


Figure B.5: Posterior distributions for the multiplicative calibration parameters for each member of the ensemble i.e, b_j, for $j = 1, \ldots, 5$. The gray line represents the true value.

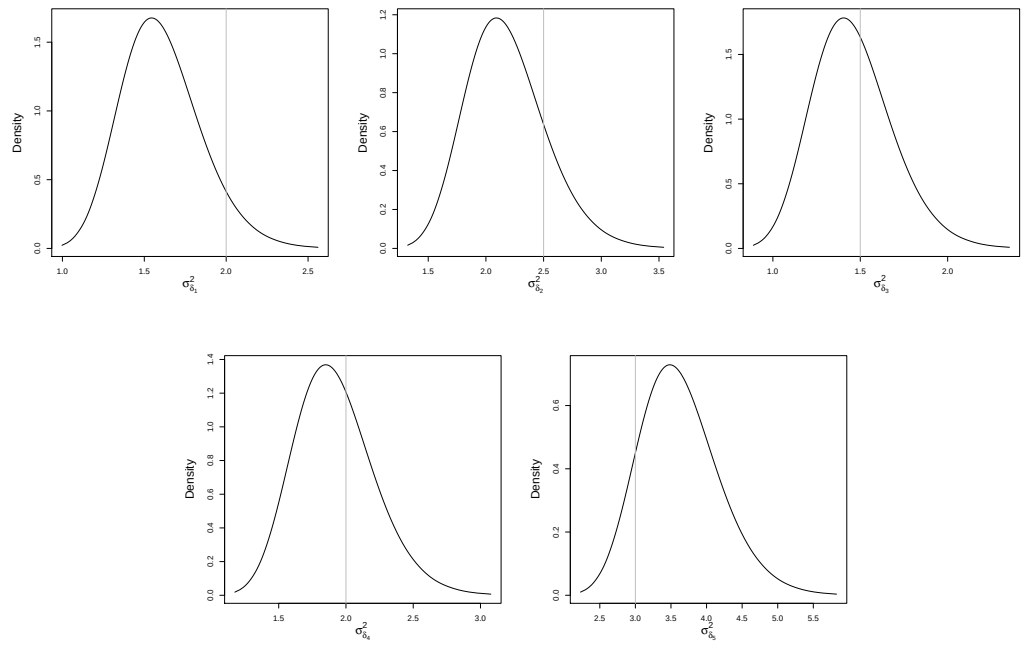


Figure B.6: Posterior distributions for the variance parameters for each member of the ensemble i.e, $\sigma^2_{\delta_j}$, for $j = 1, \ldots, 5$. The gray line represents the true value.

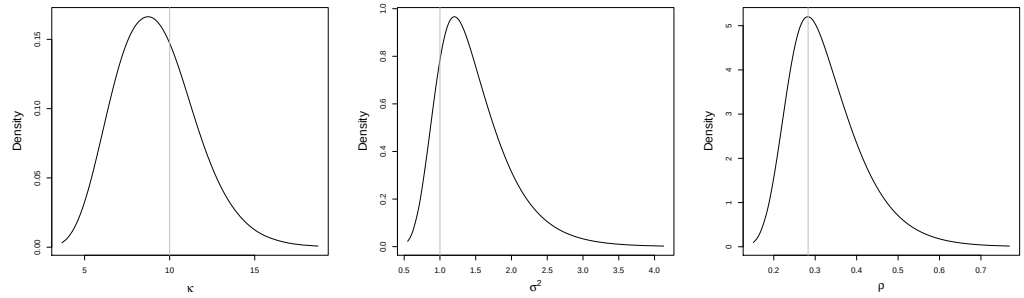


Figure B.7: Posterior distributions for covariance parameters, namely, smoothness, variance and range, respectively. The gray line represents the true value.